

MARIANNA MENDOZA HERNÁNDEZ

**ANÁLISE GEOESTATÍSTICA E MULTIVARIADA PARA DEFINIÇÃO DE
ZONAS DE MANEJO DE CANA-DE-AÇÚCAR (*Saccharum officinarum*) NA
GUATEMALA**

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

Orientador: Antônio Policarpo Souza Carneiro

Coorientadores: Gérson Rodrigues dos Santos
Paulo Cesar Emiliano

**VIÇOSA - MINAS GERAIS
2021**

**Ficha catalográfica elaborada pela Biblioteca Central da Universidade
Federal de Viçosa - Campus Viçosa**

T

M539a
2021
Mendoza Hernández, Marianna, 1991-
Análise geoestatística e multivariada para definição de
zonas de manejo de cana-de-açúcar (*Saccharum officinarum*) na
Guatemala / Marianna Mendoza Hernández. – Viçosa, MG,
2021.

1 dissertação eletrônica (67 f.): il. (algumas color.).

Inclui anexo.

Inclui apêndice.

Orientador: Antônio Policarpo Souza Carneiro.

Dissertação (mestrado) - Universidade Federal de Viçosa.

Referências bibliográficas: f. 59-62.

DOI: <https://doi.org/10.47328/ufvbbt.2021.079>

Modo de acesso: World Wide Web.

1. Agricultura de precisão. 2. Análise por agrupamento.
3. Química do solo. 4. Física do solo. 5. Análise de componentes
principais. 6. Krigagem. I. Universidade Federal de Viçosa.
Departamento de Estatística. Programa de Pós-Graduação em
Estatística Aplicada e Biometria. II. Título.

CDD 22. ed. 631.3

MARIANNA MENDOZA HERNÁNDEZ

**ANÁLISE GEOESTATÍSTICA E MULTIVARIADA PARA DEFINIÇÃO DE ZONAS
DE MANEJO DE CANA-DE-AÇÚCAR (*Saccharum officinarum*) NA GUATEMALA**

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

APROVADA: 28 de junho de 2021.

Assentimento:



Marianna Mendoza Hernández
Autora



Antônio Policarpo Souza Carneiro
Orientador

A Deus;
A minha mãe, Elsa Regina Hernández Márquez;
Ao meu pai, Bernardo Samuel Mendoza Cruz.
DEDICO.

AGRADECIMENTOS

A Deus, por me abençoar com esta grande conquista, pela sabedoria em meus estudos, por me guiar, cuidar de mim, me proteger e por ter colocado gente boa no meu caminho.

Ao Programa de Alianças para a Educação e Capacitação, Organização dos Estados Americanos e Grupo de Cooperação Internacional de Universidades Brasileiras (Brasil PAEC OEA-GCUB) e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pela oportunidade de permitir meus estudos da pós-graduação com a concessão da bolsa de mestrado.

Ao engenheiro Marino Barrientos García, por ser o professor que despertou em mim o interesse pelos estudos em Estatística, pela confiança no meu potencial e pela recomendação para eu me candidatar à bolsa de mestrado.

À Universidade Federal de Viçosa (UFV) e ao Programa de Pós-graduação em Estatística Aplicada e Biometria (PPESTBIO) pela oportunidade de realizar meus estudos de mestrado.

Ao meu orientador, professor Antônio Policarpo Souza Carneiro, e meus coorientadores, professores Gérson Rodrigues dos Santos e Paulo Cesar Emiliano, pela comunicação eficiente, por seus conhecimentos compartilhados, por suas contribuições, sugestões e orientações nesse trabalho.

Aos professores do departamento de Estatística (DET), pelos ensinamentos transmitidos.

Ao Auxiliar Administrativo do PPESTBIO, Júnior José Pires pela empatia, pela gentileza e apoio em todos os trâmites acadêmicos e administrativos durante o meu mestrado.

À empresa “*Ingenio La Unión*” e a Alvaro Rodriguez, por disponibilizarem e detalharem o banco de dados utilizado neste trabalho.

A minha mãe, Elsa Regina Hernández Márquez por sua grande confiança, amor incondicional, atenção, conselhos e por cada uma de suas palavras cheias de inspiração. Obrigada por sempre pedir a Deus por mim nesta etapa.

Ao meu pai, Bernardo Samuel Mendoza Cruz, por sua ajuda, por acreditar em mim e me motivar a dar o melhor de mim.

Aos meus irmãos Ana Lucía, Regina e Samuel, por todo apoio, motivação e confiança. Aos meus sobrinhos Katherine, Ilde Jr. e Ricardo, pelas expressões de amor que me inspiraram a cumprir esta meta.

A Keyla Patzán, pelo impulso, apoio, chamadas e mensagens de motivação, por suas orações e bons desejos para mim em todos os momentos. Por sempre me dizer: *"vos podés, con todo, Marianna"*.

Aos meus amigos, Paola Ochoa, Milvia Liquez, Marvin Pec, Ingrid Avila, Ligia Monzón, Migdalia Navarro, ao senhor Andrés Rosado, à senhora Irma Rosales e à Licenciada Heidi Grajeda, pelo carinho, por me motivarem e por confiarem em mim.

Aos meus colegas do PPESTBIO, Adenis Silva, Jhenniffer Nascimento, Gleyner de Menezes, Sabrina de Cristo, Wagner Barbosa, George de Moura, Momate Ossifo, Izabela Fialho, Calisto Máquina e Zeferino Neto, pela paciência, apoio, troca de conhecimentos e momentos de estudo.

Ao grupo de embaixadores da UFV, com menção especial a Paloma Fagundes e Carol Campos, pela ajuda, amizade, convivência e troca de experiências culturais.

À Diretoria de Relações Internacionais da UFV, em especial à Elissa Müller, técnica em assuntos educacionais, pelo apoio e empatia presencialmente e em linha nos processos de estrangeiros acadêmicos e administrativos.

RESUMO

HERNÁNDEZ, Marianna Mendoza, M.Sc., Universidade Federal de Viçosa, junho de 2021. **Análise Geoestatística e multivariada para definição de zonas de manejo de cana-de-açúcar (*Saccharum officinarum*) na Guatemala.** Orientador: Antônio Policarpo Souza Carneiro. Coorientadores: Gérson Rodrigues dos Santos e Paulo Cesar Emiliano.

As zonas de manejo (ZM) são delimitações de subáreas com características comuns. Dentro da mesma ZM, o potencial de produção pode ser semelhante, ao passo que pode diferir entre as diversas ZM. Estudos recentes em agricultura de precisão consideram eficiente o uso de ZM para a aplicação de fertilizantes baseada na variabilidade espacial dos atributos do solo. Diante disso, os objetivos deste trabalho foram analisar a variabilidade espacial dos atributos físico-químicos do solo em lavoura de cana-de-açúcar (*Saccharum officinarum*), por meio de Geoestatística e delinear zonas de manejo homogêneas usando análise Geoestatística e multivariada. Foram analisados dados de uma área de produção de cana-de-açúcar de 1.516,33 ha, localizada na fazenda Tehuantepec, no município de *Santa Lucía Cotzumalguapa, Escuintla*, Guatemala. Foram utilizados 17 atributos físico-químicos, obtidos da amostragem do solo de 153 pontos georreferenciados da área de estudo, na safra 2019-2020. Foram eles: potencial hidrogeniônico, fósforo, cobre, zinco, ferro, manganês, boro, matéria orgânica, capacidade de troca catiônica, cálcio, magnésio, sódio, potássio, saturação de bases, argila, silte e areia. A determinação da dependência espacial de cada variável foi analisada pela análise variográfica, considerando para seu ajuste as variáveis que não apresentaram efeito pepita puro. Os modelos esférico e exponencial foram ajustados aos variogramas experimentais usando o método dos mínimos quadrados. Para avaliar a qualidade do ajuste do modelo teórico dos variogramas, utilizou-se o processo de validação cruzada e, para delimitar as zonas de manejo, foi utilizado o algoritmo fuzzy K-means. A análise de componentes principais (ACP) foi aplicada aos valores preditos dos atributos do solo, obtidos pela técnica de interpolação krigagem como método de entrada de dados para o algoritmo fuzzy K-means. O número de ZM foi definido com os critérios do coeficiente de partição difusa (FPC) e entropia de partição normalizada (NCE). A área estudada apresentou variabilidade espacial para 16 dos 17 atributos de solo. Foram delineadas duas zonas de manejo para a cultura da cana-de-açúcar.

Palavras-chave: Agricultura de precisão. Atributos físico-químicos do solo. Componentes principais. Krigagem. Variabilidade espacial. Fuzzy K-means.

ABSTRACT

HERNÁNDEZ, Marianna Mendoza, M.Sc., Universidade Federal de Viçosa, June, 2021. **Geostatistical and multivariate analysis for defining management zones for the cultivation of sugarcane (*Saccharum officinarum*) in Guatemala.** Adviser: Antônio Policarpo Souza Carneiro. Co-advisers: Gérson Rodrigues dos Santos and Paulo Cesar Emiliano.

The management zones (MZs) are a delimitation of subareas with characteristics in common. Within the same management zone, the production potential may be similar, while it may differ between the other MZs. Recent studies in precision agriculture consider efficient the use of MZs for the application of fertilizers based on the spatial variability of the soil attributes. The objectives of this work were to analyze the spatial variability of the physical-chemical attributes of the soil in the cultivation of sugar cane (*Saccharum officinarum*), by means of Geostatistics and to delineate homogeneous management areas using Geostatistical analysis and multivariate analysis. Data from a sugarcane production area of 1,516.33 hectares, located in the Tehuantepec farm, municipality of Santa Lucía Cotzumalguapa, Escuintla, Guatemala were analyzed. 17 physical-chemical attributes were used, obtained from soil sampling of 153 georeferenced points in the study area, corresponding to the 2019-2020 harvest. The soil variables were: potential for hydrogen, phosphorus, copper, zinc, iron, manganese, boron, organic matter, cation exchange capacity, calcium, magnesium, sodium, potassium, base saturation, clay, silt and sand. The determination of spatial dependence of each variable was determined by variographic analysis, considering for its adjustment the variables that did not present a pure nugget effect. The spherical and exponential models were fitted to the experimental variograms using the least squares method. To evaluate the quality of fit of the theoretical model of the variograms, the cross-validation process was used and, to delimit the management areas, the Fuzzy K-means algorithm was used. The principal component analysis (PCA) was applied to the predicted values of the soil attributes, obtained by the kriging interpolation technique as the data entry method for the Fuzzy K-means algorithm. The number of management zones was defined with the criteria of diffuse partition coefficients (FPC) and normalized partition entropy (NCE). The study area presented special variability for 16 of the 17 soil attributes. Two management zones for the cultivation of sugar cane were delineated.

Keywords: Precision agriculture. Management zones. Physicochemical soil attributes. Component principal analysis. Kriging. Spatial variability. Fuzzy K-means.

LISTA DE ILUSTRAÇÕES

Figura 1 - Parâmetros do variograma experimental e variograma ajustado por um modelo teórico obtido a partir das semivariâncias estimadas γh em função das distâncias h	19
Figura 2 - Mapa de localização do município Santa Lucía Cotzumalguapa, Escuintla, Guatemala.....	30
Figura 3 - Campo amostral dos 153 pontos de amostragem do solo georreferenciados na fazenda produtora de cana-de-açúcar Tehuantepec, localizada na região sul da Guatemala, no município de Santa Lucía Cotzumalguapa, Escuintla.	31
Figura 4 - Matriz de correlação linear de Pearson para os atributos do solo da área de estudo de produção de cana-de-açúcar.	36
Figura 5 - Variograma empírico e variograma teórico ajustado pelo modelo esférico para os atributos potencial hidrogeniônico, fósforo, cobre, zinco, ferro e manganês.....	38
Figura 6 - Variograma empírico e variograma teórico ajustado pelo modelo esférico para os atributos boro, matéria orgânica, capacidade de troca catiônica, cálcio, magnésio e modelo exponencial para a variável sódio.....	39
Figura 7 - Variograma empírico e variograma teórico ajustado pelo modelo esférico para os atributos potássio, saturação de bases, argila, silte e areia.	40
Figura 8 - Mapas de distribuição espacial obtidos a partir da krigagem para os atributos: potencial hidrogeniônico, fósforo, cobre, zinco, ferro e manganês.....	43
Figura 9 - Mapas de distribuição espacial obtidos a partir da krigagem para os atributos: boro, matéria orgânica, capacidade de troca catiônica, cálcio, magnésio e sódio.	44
Figura 10 - Mapas de distribuição espacial obtidos a partir da krigagem dos atributos: potássio, saturação de bases, argila e areia.	45
Figura 11 - Gráfico de Scree plot dos autovalores em função do número de componentes principais das 16 variáveis do solo que apresentaram dependência espacial.	46

Figura 12 - Mapas de distribuição espacial do primeiro componente principal e das primeiras quatro variáveis mais correlacionadas: fósforo, saturação de bases, manganês e matéria orgânica.	49
Figura 13 - Mapas de distribuição espacial do componente principal 2 e das primeiras quatro variáveis mais correlacionadas: ferro, fósforo, capacidade de troca catiônica e matéria orgânica.	50
Figura 14 - Mapas de distribuição espacial do componente principal 3 com as variáveis mais correlacionadas, com coeficientes maiores a 0,50: fósforo, zinco e saturação de bases.	50
Figura 15 - Mapas de distribuição espacial do componente principal 4 com as variáveis mais correlacionadas, com coeficientes maiores a 0,50 em módulo: fósforo, saturação de bases, potencial hidrogeniônico e matéria orgânica.	50
Figura 16 - Índices de validação de zonas de manejo em relação ao número de cluster da área de estudo para duas zonas de gerenciamento.	51
Figura 17 - Mapa da delimitação de duas zonas de manejo para os 2 cluster obtidos a partir do algoritmo Fuzzy K-means.	52
Figura 18 - Análise exploratória gráfica dos atributos do solo. a) Potencial hidrogeniônico, b) fósforo, c) cobre, d) zinco, e) ferro, f) manganês, g) boro, h) matéria orgânica.	63
Figura 19 - Análise exploratória gráfica dos atributos do solo. a) Capacidade de troca catiônica, b) cálcio, c) magnésio, d) sódio, e) potássio, f) saturação de bases, g) argila, h) silte.	64
Figura 20 - Análise exploratória gráfica do atributo areia.....	65
Figura 21 - Incremento da produção em toneladas de cana por hectare (THC, %), relacionado ao conteúdo de matéria orgânica expressada em porcentagem (MO, %), como resposta às aplicações de nitrogênio na cultura de cana-de-açúcar na Guatemala.	67

LISTA DE TABELAS

Tabela 1 - Modelos teóricos de variogramas. Funções de variograma para o modelo esférico, exponencial e gaussiano.	21
Tabela 2 - Estatística descritiva para os atributos físicos e químicos do solo da área produtora de cana-de-açúcar com medições georreferenciadas em 153 locais amostrados.....	35
Tabela 3 - Parâmetros variográficos e resultados da validação cruzada dos valores preditos pela técnica de interpolação Krigagem e valores observados dos atributos do solo.....	41
Tabela 4 - Resultados da análise de componentes principais aplicada aos dados interpolados pela técnica Krigagem dos atributos do solo com variabilidade espacial.	46
Tabela 5 - Resultados dos autovalores, variação e variação acumulada das primeiras quatro componentes principais depois de eliminar a variável sódio (Na).	47
Tabela 6 - Autovetores e correlações entre cada componente principal e os atributos do solo, resultados obtidos a partir da ACP.	49
Tabela 7 - Coeficiente de partição difusa (FPC) e entropia de partição normalizada (NCE) em relação ao número de <i>cluster</i> da área de estudo, com base em quatro componentes principais, calculados a partir dos atributos do solo.....	51
Tabela 8 - Medidas descritivas dos dados gerados para duas zonas de manejo delimitadas pelo algoritmo Fuzzy K-means.	53
Tabela 9 - Comparação entre os valores da média aritmética, desvio padrão e coeficiente de variação para a área total e para as duas zonas de manejo delimitadas pelo algoritmo Fuzzy K-means dos atributos do solo em estudo.	54
Tabela 10 - Amplitude de valores adequados dos atributos dos solos característicos das áreas de produção de cana-de-açúcar da região sul da Guatemala segundo Pérez (2013), com conteúdo menor a 35 % de argila e valores dos atributos georreferenciados para as duas zonas de manejo.....	55
Tabela 11 - Matriz de correlação linear de Pearson.	66

SUMÁRIO

1	INTRODUÇÃO.....	12
2	REVISÃO BIBLIOGRÁFICA.....	14
2.1	Agricultura de precisão.....	14
2.2	Zonas de manejo.....	15
2.3	Geoestatística.....	16
2.3.1	Krigagem.....	22
2.4	Análise de componentes principais (ACP).....	23
2.4.1	Número de componentes principais.....	24
2.4.1.1	Número de variáveis.....	25
2.4.2	Algoritmos de <i>Clustering</i>	26
2.4.2.1	K-means.....	26
2.4.2.2	Algoritmo Fuzzy K-means.....	27
2.4.3	Índices de validação de ZM.....	28
2.4.4	Análise de variância (ANOVA).....	29
3	MATERIAL E MÉTODOS.....	30
3.1	Descrição da área.....	30
3.2	Estatística descritiva.....	31
3.3	Análise Geoestatística.....	31
3.4	Análise de componentes principais.....	32
3.5	Análise de agrupamento para delinear zonas de manejo.....	32
4	RESULTADOS E DISCUSSÃO.....	33
4.1	Estatística descritiva.....	33
4.2	Análise de correlação.....	36
4.3	Análise Geoestatística.....	37
4.4	Análise de componentes principais (ACP).....	45
4.5	Análise de agrupamento – Algoritmo Fuzzy K-means.....	51
5	CONCLUSÃO.....	58
	REFERÊNCIAS BIBLIOGRÁFICAS.....	59
	APÊNDICE.....	63
	ANEXO.....	67

1 INTRODUÇÃO

A agricultura específica para cada local precisa da identificação de regiões, denominadas zonas de manejo (ZM). As ZM são delimitações de subáreas dentro do campo, nas quais cada ZM possui características comuns. As delimitações de ZM também são úteis para orientar a determinação do número de amostras para análise de solo e de plantas (MORAL; TERRÓN; SILVA, 2010).

A variabilidade espacial da produtividade em determinada cultura é afetada por vários fatores, como propriedades físico-químicas e orgânicas do solo, fazendo-se necessário detectar os fatores mais relevantes por meio da análise Geoestatística (ORTEGA; SANTIBÁÑEZ, 2007). Estudos já foram realizados demonstrando como os atributos do solo fornecem uma base técnica na definição de ZM no cultivo de arroz (FU; WANG; JIANG, 2010), soja, milho, trigo e aveia preta (MOLIN; DE CASTRO, 2008), algodão (LI et al., 2007), dendezeiros (BEHERA et al., 2018), uva (OHANA-LEVI et al., 2019) e cana-de-açúcar (SOUZA et al., 2010).

A Geoestatística permite a análise e modelagem de variabilidade espacial e temporal de um conjunto de dados georreferenciados, obtendo modelos de dependência espacial por meio de variogramas e seus ajustes, a interpolação dos dados pelo método krigagem e a geração de mapas (BERNARDI et al., 2014). As etapas da Geoestatística são: i) análise variográfica, ii) validação dos modelos teóricos e iii) predição em locais não amostrados usando a técnica de krigagem. Vale ressaltar que o variograma experimental corresponde a uma distância média entre pares de pontos dentro de cada intervalo de distância, e não a uma distância específica h . Portanto, é necessário um ajuste por meio de modelos teóricos que generalizem as observações do variograma experimental a qualquer distância, obtendo como resultado o variograma ajustado.

Os modelos teóricos mais amplamente utilizados são os modelos esférico, exponencial e gaussiano (VIEIRA, 2000). Para avaliar a qualidade do ajuste do modelo teórico, é necessária a realização da validação cruzada (VIEIRA, 2000; HENRÍQUEZ; MÉNDEZ; MASÍS, 2013). Em seguida, ao comparar os modelos, escolhe-se aquele com o melhor ajuste. Então, com base nos valores do variograma ajustado é realizada a etapa de krigagem. A krigagem é um método de interpolação Geoestatística usado para estimar valores de variáveis em qualquer posição a partir de valores adjacentes (ROSA, 2017 apud Landim 1998). Ela permite interpolar valores sem tendência y com variância mínima (VIEIRA, 2000).

Por sua vez, as técnicas multivariadas da análise de componentes principais (ACP) e análise de agrupamento são utilizadas para definir ZM homogêneas com base em padrões de

variabilidade espacial das culturas e atributos do solo (OLDONI et al., 2019). ACP é uma técnica de redução de dimensionalidade na qual os componentes principais devem ser não correlacionados entre si e conter a maior parte da variância total explicada do conjunto de novas variáveis. O número de componentes principais a reter na análise é baseado nos seguintes critérios: explicar pelo menos 70% da variação total, autovalores ≥ 1 e utilizar o procedimento gráfico *scree plot*. Nesse gráfico, os autovalores (eixo y) são plotados com a ordem i (eixo x) das componentes principais. O número de componentes a serem retidas é o componente k , situado no ponto no qual os autovalores tendem a se estabilizar (FERREIRA, 2018).

Os algoritmos de *clustering* são métodos de agrupamento de objetos semelhantes no mesmo grupo homogêneo, sendo dissimilares entre grupos diferentes. Os métodos mais utilizados são os algoritmos de análise cluster K-means e fuzzy K-means (GAVIOLI et al., 2019; LIKAS; VLASSIS; VERBEEK, 2020). O algoritmo K-means tem por objetivo minimizar a expressão da soma do erro quadrático em todos os grupos K. Neste procedimento, a distância euclidiana é usada como norma para expressar a distância entre as observações. Por sua vez, o Fuzzy K-means é uma extensão do método K-means, em que cada ponto dos dados pode fazer parte de vários clusters com um valor de associação, chamado atribuição suave. Esse algoritmo usa um expoente de ponderação para controlar o grau em que o compartimento de associação ocorre entre os clusters (JAIN, 2010). A classificação de ZM pelo método Fuzzy K-means tem sido utilizada para estudos de rendimentos e atributos do solo de culturas, tais como milho, soja, algodão, pessegueiro e tabaco (GAVIOLI et al., 2016; GILI et al., 2017; LI et al., 2007; OLDONI et al., 2019; XIN-ZHONG et al., 2009), e também em estudos de condutividade elétrica (capacidade que um material possui em conduzir corrente elétrica) e atributos do solo de cultivos de soja, milho e arroz (PERALTA; COSTA, 2013; TRIPATHI et al., 2015).

O presente estudo tem por objetivo geral delinear zonas de manejo (ZM) para a fazenda Tehuantepec, com produção de cana-de-açúcar na Guatemala, usando análise Geoestatística multivariada. Os objetivos específicos deste trabalho são: analisar a variabilidade espacial dos atributos do solo cultivados com cana-de-açúcar (*Saccharum officinarum*) por meio de ferramentas Geoestatística e delinear zonas de manejo homogêneas usando análise Geoestatística multivariada.

2 REVISÃO BIBLIOGRÁFICA

2.1 Agricultura de precisão

A cada ano, a agricultura enfrenta desafios de aumentar a produção para atender à demanda da população. O crescimento da produção é dado principalmente pela expansão de novas áreas agrícolas e pelo aumento da produtividade. O aumento da produtividade por unidade de área tem sido obtido pelo uso de tecnologias novas e mais eficientes, que permitem quantificar e gerenciar de forma diferenciada a variabilidade existente na área produtora (BONGIOVANNI et al., 2006).

O Sistema de Posicionamento Global (GPS) iniciou as primeiras operações em 1978 e foi lançado para uso civil no final da década de 80/início da década de 90, possibilitando o desenvolvimento de equipamentos inteligentes que permitem a obtenção de dados georreferenciados de produção e aplicação de insumos agrícolas localizados (BERNARDI et al., 2014). Com base nos sistemas e processos inovadores, começou a surgir a agricultura de precisão, considerando as características do solo e da lavoura no tempo e espaço.

Segundo Bongiovanni et al. (2006), a agricultura de precisão (AP) é o conjunto de técnicas orientadas a otimizar as aplicações dos insumos agrícolas baseadas na variabilidade espaço-temporal da produção agrícola. A identificação, medição e gestão da variabilidade espacial dos fatores relacionados aos sistemas de produção permitem o manejo específico do local por meio do delineamento de áreas com características homogêneas para apoiar a tomada de decisões associadas à produção da lavoura (INIA, 1999). Outra designação de AP é a agricultura específica do local, que engloba a coleta e análise de dados em diferentes locais dentro de uma determinada área para fornecer um tratamento específico de gestão em cada local (MORAL; TERRÓN; REBOLLO, 2011).

A variabilidade espacial do rendimento em determinada cultura é afetada por vários fatores, como propriedades físico-químicas do solo, clima, topografia, práticas de manejo, fertilização, irrigação, pragas, doenças, fisiologia das plantas e também pela interação dos fatores acima mencionados. Como a identificação de todos os fatores que limitam a produção é uma tarefa complicada, torna-se necessário detectar os fatores mais relevantes através da análise estatística (ORTEGA; SANTIBÁÑEZ, 2007). Desde o início da agricultura de precisão, várias investigações têm sido conduzidas com foco no manejo local específico denominado zonas de manejo. Assim, o rendimento potencial do ponto de vista da fertilidade do solo é

alcançado quando o cultivo tem condições ótimas de água e nutrientes na ausência de fatores limitantes (CENGICAÑA, 2012).

2.2 Zonas de manejo

As zonas de Manejo são delimitações de subáreas dentro do campo, nas quais cada ZM tem características em comum. A obtenção de áreas de manejo mais uniformes são definidas com fatores relacionados com o rendimento das culturas ao longo de tempo e espaço (LI et al., 2007).

Dados georreferenciados sobre rendimento da cultura, propriedades do solo, fatores topográficos, índices de vegetação e possíveis combinações podem ser avaliados para definir zonas de manejo (GAVIOLI et al., 2016). Por exemplo, em um estudo, o delineamento de ZM de nutrientes do solo foi capaz de fornecer a base técnica para planos de fertilização de arroz entre subzonas específicas na AP (FU; WANG; JIANG, 2010).

Em seu trabalho, Molin e de Castro (2008) verificaram que é possível delimitar zonas de manejo diferenciadas usando dados de condutividade elétrica (CE) do solo (de leitura superficial e profunda) e propriedades físico-químicas (fósforo, potássio, cálcio, magnésio, saturação de base, capacidade de troca de cátions, soma de bases, matéria orgânica, pH, teor de argila e teor de areia). O estudo foi realizado em uma área com lavouras de soja, milho, trigo e aveia preta. Dados de CE aparente medido em suspensão, CE aparente medido no modo de superfície e CE aparente medido no modo profundo e dados de propriedades físicas do solo também foram considerados para delinear ZM numa plantação de *Brassica napus* L. (MORAL; TERRÓN; SILVA, 2010).

Por sua vez, Li et al. (2007) definiram zonas de manejo em região salina para cultivo de algodão. Os autores avaliaram a relação do rendimento do algodão com dados de propriedades físico-químicas do solo e dados do índice de vegetação por diferença normalizada (NDVI). Os resultados apontaram que as informações utilizadas são viáveis para a tomada de decisão de maneira racional e definida por local. Além disso, a divisão em ZM demonstrou ser útil para o desenvolvimento de um plano específico de amostragem do solo, que serve para determinar a variabilidade do rendimento das culturas com base nas propriedades do solo. A divisão também permitiu reduzir o número de amostras para análises de solo necessárias para a criação de mapas operacionais da cultura, contribuindo com a redução do custo para caracterizar a variabilidade espacial do campo. Em outro estudo, Behera et al. (2018) avaliaram a distribuição espacial dos atributos do solo e o delineamento de ZM em plantações de dendezeiros para recomendação

eficiente de fertilização. A análise de agrupamento em conjunto com a caracterização da variabilidade espacial dos macronutrientes do solo foi eficaz na definição de ZM para aplicação de insumos agrícolas.

Gavioli et al. (2016) estudaram a otimização de zonas de manejo em três áreas agrícolas de milho e soja, com dados dos anos de 2010 a 2014. Para o estudo, foram utilizadas as variáveis: resistência à penetração no solo (RPS) a 0-0,1 m, RPS a 0,1-0,2, RPS a 0,2-0,3 m, pH, elevação, declive, densidade, teor de areia, teor de silte, teor de argila, teor de matéria orgânica e dados de produtividade de soja e milho. Dos seis métodos de seleção de variáveis utilizados para delinear as ZM, a análise espacial multivariada com base no índice ACP de Moran mostrou ser a melhor. Ela foi aplicada a variáveis significativamente correlacionadas com o rendimento de cada área, que apresentou maior homogeneidade nas áreas de manejo, tornando-o mais viável para sua implementação operacional nas terras agrícolas.

Ohana-Levi et al. (2019) realizaram um estudo sobre delineamento de ZM de irrigação em campos de produção de uva. Os autores verificaram a viabilidade de representar espacialmente um grande conjunto de dados e exploraram o comportamento espacial multivariado em culturas agrícolas. Para isso, estudaram dados de produção e doze variáveis preditivas (índice de estresse hídrico, índice de vegetação com diferença normalizada (NDVI), níveis de cloro, condutividade elétrica, acidez, condutividade elétrica aparente, porcentagem de saturação, profundidade do solo, aspecto, inclinação, índice de umidade topográfica, distância das árvores circundantes). O NDVI foi a variável mais dominante e influente na explicação da variabilidade espacial do rendimento. O uso de grande variedade de recursos de dados forneceu boa delimitação de ZM de irrigação.

2.3 Geoestatística

A estatística espacial é uma área da estatística que estuda processos aleatórios, considerando a distribuição geográfica de variáveis aleatórias. O objetivo da estatística espacial é a coleta, exploração, descrição, visualização e análise de dados espaciais obtidos de um grupo de observações feitas em pontos georreferenciados, com coordenadas de latitude e longitude. A estatística espacial pode ser classificada em três grandes áreas: padrões espaciais, dados *lattice* e geoestatística. A geoestatística se diferencia pelo fato de trabalhar com dados espacialmente contínuos e contempla a modelagem do padrão de variabilidade espacial e predição em pontos não amostrados (LOUZADA, 2011).

A geoestatística teve seu início na África do Sul nas décadas de 50 e 60, quando Daniel Gerhardus Krige, trabalhando com dados de reservas de concentração de ouro, percebeu que era preciso considerar a distância e a direção da amostragem para realizar estimações nas minas (SANTOS, 2010). Georges Matheron desenvolveu os fundamentos geoestatísticos baseados na ideia de Krige, que leva em consideração a localização geográfica por meio da publicação da “Teoria das Variáveis Regionalizadas”. Esta teoria adota uma perspectiva estocástica dos processos espaciais, com uma função numérica que varia de um local para outro, com uma continuidade ou dependência espacial (VIEIRA, 2000).

A teoria das variáveis regionalizadas tem por pressuposição que a variação dos dados é expressa pela soma da componente estrutural, a componente aleatória com correlação espacial e o ruído aleatório não correlacionado. Sendo $Z(x)$ uma variável regionalizada, sua expressão matemática em função das componentes pode ser denotada por:

$$Z(x) = \mu(x) + \varepsilon'(x) + \varepsilon''$$

Segundo Matheron (1970), para fazer inferência estatística é preciso introduzir hipóteses para reduzir o número de parâmetros dos quais depende a função aleatória $Z(x)$. Sob esse objetivo, o autor definiu as hipóteses de estacionariedade de 1ª ordem, 2ª ordem e intrínseca. Uma função aleatória estacionária se repete de certa maneira no espaço permitindo que seja possível realizar a inferência estatística com apenas uma repetição. Isso significa que os momentos estatísticos da variável aleatória $Z(x + h)$ são os mesmos para qualquer distância denominada h (SANTOS, 2010).

Para que a função aleatória $Z(x)$, seja estacionária de 1ª ordem, as seguintes características devem ser satisfeitas para qualquer distância h : $E[Z(x)] = \mu$ e $\text{var}[Z(x)] = \sigma^2$. Por sua vez, a estacionariedade de 2ª ordem é assumida quando as variáveis aleatórias $Z(x)$ e $Z(x + h)$ que modelam os valores verdadeiros dos dados têm o mesmo valor esperado e que a covariância entre duas dessas variáveis não depende de suas localizações específicas, mas apenas do vetor de separação entre elas. A expressão matemática é dada por $E[Z(x)] = \mu$ e $E[Z(x + h) - Z(x)] - \mu^2 = \text{cov}(h)$. Sob esta suposição, a relação entre o variograma e o covariograma é dada por $\gamma(h) = \text{cov}(0) - \text{cov}(h) = \text{var}[Z(x)] - \text{cov}(h)$.

Segundo Matheron (1970), para o caso de não existir uma variância finita é necessário substituir o suposto de estacionariedade de 2ª ordem pela hipótese intrínseca. A função aleatória $Z(x)$ é intrínseca se para todo vetor h o incremento $Z(x + h) - Z(x)$ tem o valor esperado e variância que não depende de x , o que é expresso matematicamente como:

$$E[Z(x+h) - Z(x)] = \mu$$

$$E[Z(x+h) - Z(x)]^2 = 2\gamma(h)$$

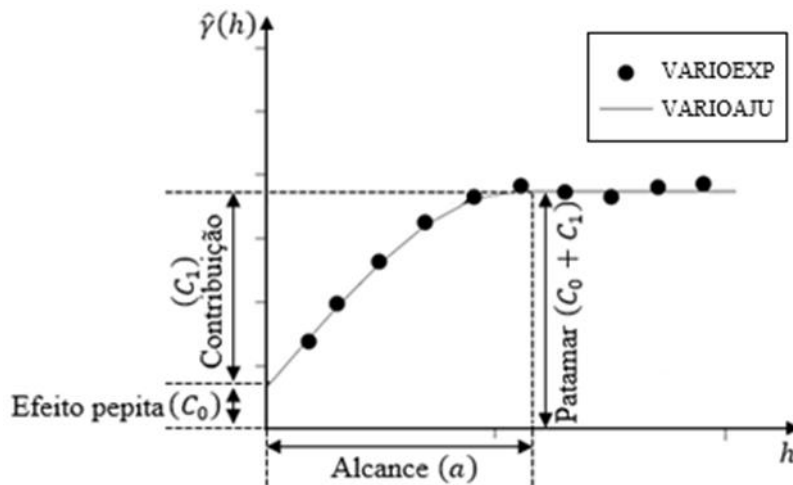
Os principais objetivos da teoria das variáveis regionalizadas são dois: o primeiro é a análise estrutural ou variográfica e o segundo é a predição em locais não amostrados usando a técnica de krigagem. Levando em consideração tais objetivos, o erro de estimação depende das características variográficas (MATHERON, 1970). A análise variográfica descreve quantitativamente a variabilidade espacial entre pontos vizinhos. A dependência espacial existe quando a semivariância apresenta pequenos valores para pares de pontos localizados muito próximos um do outro. Quanto mais perto dois pontos estão localizados, mais são correlacionados (ROSA, 2017).

O valor da semivariância para qualquer distância h pode ser calculado com a estimativa $\hat{\gamma}(h)$, considerando $N(h)$ como o número de pares de pontos separados entre si por uma distância h ; $z(x_i)$ como valor observado da variável no ponto (x_i) ; e $z(x_i + h)$ como valor observado no ponto $(x_i + h)$. O estimador das semivariâncias é definido por:

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [z(x_i) - z(x_i + h)]^2$$

A partir do gráfico das estimativas das semivariâncias $\hat{\gamma}(h)$ em função das distâncias é obtido o variograma experimental. O esperado é que pontos próximos sejam mais semelhantes entre eles do que amostras mais afastadas (SANTOS et al. 2011 apud ROSA, 2017) e que as diferenças $z(x_i) - z(x_i + h)$ decresçam à medida em que diminua a distância h que os separa (VIEIRA, 2000). Os parâmetros do variograma (alcance, patamar, efeito pepita e contribuição) podem ser observados na Figura 1.

Figura 1 - Parâmetros do variograma experimental e variograma ajustado por um modelo teórico obtido a partir das semivariâncias estimadas $\hat{\gamma}(h)$ em função das distâncias h .



FONTE: Adaptado de Heano (2002).

O patamar ($C_0 + C_1$) é o valor da semivariância na qual o variograma se estabiliza e é aproximadamente igual à variância dos dados. O alcance (a) é a distância dentro da qual as amostras mostram dependência espacial. O efeito pepita (C_0) é a descontinuidade na origem, ou seja, o valor da semivariância para distâncias próximas a zero. A contribuição ou componente estrutural (C_1), por sua vez, é a diferença entre o patamar e o efeito pepita (ROSA, 2017).

Existem variogramas com patamar (considerando também o patamar assintótico) e variogramas sem patamar (com dispersão infinita). Uma variável apresenta ausência de dependência espacial quando tem efeito pepita puro, o que ocorre quando o variograma é constante e igual ao patamar para qualquer valor de distância h (VIEIRA, 2000). Segundo Cambardella et al. (1994), a classificação de dependência espacial com dados do solo pode ser determinada pela relação em porcentagem do efeito pepita em respeito ao patamar, sendo forte para valores menores que 25%, moderada entre 25% e 75% e fraca para valores maiores que 75%.

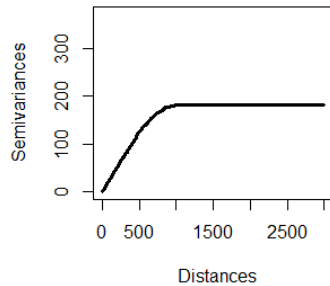
Quanto aos tipos de variogramas, existem três: variograma real ou direto, variograma experimental ou empírico e o variograma ajustado. O variograma direto consiste no estudo dos pontos de uma variável regionalizada para descrever o comportamento de dependência espacial. Já o variograma experimental é obtido pelas semivariâncias a uma distância média entre pares de pontos dentro de cada intervalo de distância dos dados da variável. Por fim, o variograma ajustado é aquele que considera o variograma experimental para seu ajuste por meio de modelos teóricos, com o objetivo de generalizar as observações do variograma empírico a qualquer

distância (BALBOA, 2012). O variograma experimental corresponde a uma distância média entre pares de pontos dentro de cada intervalo de distância, e não a uma distância específica h . Portanto, é necessário um ajuste por meio de modelos teóricos que generalizem as observações do variograma experimental a qualquer distância, modelos condicionalmente positivos definidos. Os modelos teóricos com patamar mais amplamente utilizados são os modelos esférico, exponencial e gaussiano (VIEIRA, 2000).

O modelo esférico tem comportamento linear próximo à origem, os incrementos marginais diminuem nas grandes distâncias, até que se tornem nulos (em distâncias maiores que o alcance) e seja atingido o patamar no valor do alcance (CÓRDOBA, 2014; VIEIRA, 2000). Os modelos exponencial e gaussiano atingem o patamar do variograma assintoticamente, isto é, o alcance é obtido tomando-se 95% do patamar (LOUZADA, 2011). Na Tabela 1 estão apresentados exemplos de variogramas e as expressões matemáticas dos modelos teóricos. Nestes exemplos, os parâmetros variográficos para um mesmo conjunto de dados são: alcance com 1141, efeito pepita zero e patamar 182.

Tabela 1 - Modelos teóricos de variogramas. Funções de variograma para o modelo esférico, exponencial e gaussiano.

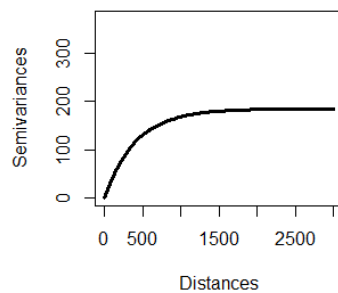
Modelo Esférico



$$\gamma(h) = C_0 + C_1 \left[\frac{3}{2} \left(\frac{h}{a} \right) - \frac{1}{2} \left(\frac{h}{a} \right)^3 \right]; \quad 0 < h < a$$

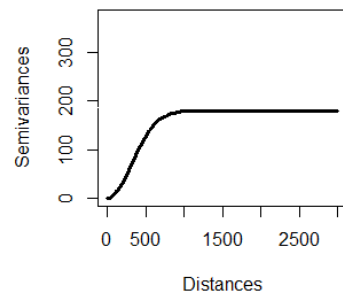
$$C_0 + C_1; \quad h \geq a$$

Modelo Exponencial



$$\gamma(h) = C_0 + C_1 \left[1 - \exp \left(-3 \frac{h}{a} \right) \right]; \quad 0 < h < d$$

Modelo Gaussiano



$$\gamma(h) = C_0 + C_1 \left[1 - \exp \left(-3 \left(\frac{h}{a} \right)^2 \right) \right]; \quad 0 < h < d$$

Fonte: Os gráficos são de elaboração própria e as equações são de VIEIRA, 2000.

Para avaliar a qualidade do ajuste do modelo teórico, é realizado o processo de validação cruzada. Este processo consiste em retirar um dado do grupo original para fazer uma predição do valor da variável na localização do dado removido e repetir o processo para todos os dados dos pontos observados com os valores estimados (HENRÍQUEZ; MÉNDEZ; MASÍS, 2013). Segundo Santos (2010), com o conjunto de dados dos valores medidos e preditos pode-se estimar uma regressão linear que permita avaliar a qualidade da predição.

$$\hat{Z}(x_i) = a + bZ(x_i)$$

Em que:

- $Z(x_i)$: é o valor medido em cada observação x_i ;
- $\hat{Z}(x_i)$: é o valor estimado em cada observação x_i ;
- a : é o intercepto, que preferencialmente deve ser zero. Valores mais afastados do zero indicam superestimação (+) ou subestimação (-);
- b : é o coeficiente angular da reta da regressão, que preferencialmente deve ser igual a 1.

Para avaliar a qualidade da predição de ajuste no variograma, realiza-se a modelagem de variogramas empíricos com dois ou mais modelos e, para cada modelo, estima-se $Z(x_i)$ e a variância da krigagem $\hat{\sigma}^2(x_i)$ é obtida para o cálculo do erro médio (ME), erro quadrático médio (MSE) e erro padrão (SE). Os erros obtidos a partir da diferença entre os valores observados e os valores preditos de um bom modelo tendem a uma média aproximadamente zero e variância constante. Ao padronizar os erros, a média deve ser aproximadamente 0 e a variância 1 (WEBSTER & OLIVER, 2007).

$$ME = \frac{1}{N} \sum_{i=1}^N [Z(x_i) - \hat{Z}(x_i)]$$

$$MSE = \frac{1}{N} \sum_{i=1}^N [Z(x_i) - \hat{Z}(x_i)]^2$$

$$SE = \frac{1}{N} \sum_{i=1}^N \frac{[Z(x_i) - \hat{Z}(x_i)]^2}{\hat{\sigma}^2(x_i)}$$

2.3.1 Krigagem

Ao comparar-se os modelos, deve ser escolhido aquele com o melhor ajuste. Com base nos valores do variograma ajustado, é realizada a etapa de krigagem. A krigagem é um método de interpolação Geoestatística, usada para estimar valores de variáveis em qualquer posição a partir de valores adjacentes (LANDIM, 1998 apud ROSA, 2017). Ela permite interpolar valores atribuindo pesos a cada amostra por meio de preditores que usam a vizinhança amostrada com dependência espacial para reduzir a incerteza das predições, obtendo uma mínima variância de estimação (MATHERON, 1970; SANTOS, 2010). A krigagem de tipo linear calcula as estimativas de variáveis aleatórias por meio de combinações lineares dos dados observados,

visando interpolar valores sem tendência e com variância mínima (VIEIRA, 2000). Os métodos lineares da krigagem mais utilizados são: krigagem simples, ordinária e universal.

A krigagem simples supõe a função média conhecida para a variável regionalizada. A krigagem ordinária (KO) assume o valor esperado estacionário, mas desconhecido, considerando apenas o variograma, por isso é a krigagem mais utilizada. A krigagem universal, por sua vez, não satisfaz a condição de estacionariedade da variável regionalizada e é caracterizado por exibir uma tendência. Dentre os preditores lineares da geoestatística, a krigagem simples é o método mais preciso, seguido pela ordinária e pela universal (SANTOS, 2010).

2.4 Análise de componentes principais (ACP)

A análise de componentes principais (ACP) é uma técnica multivariada de redução de dimensionalidade que visa, a partir de grande número de variáveis originais relacionadas, obter um número menor de variáveis latentes. Essas variáveis latentes são denominadas componentes principais, que são combinações lineares das variáveis originais. A ACP baseia-se na realização de uma rotação rígida no sistema de eixos coordenados das variáveis originais a novos eixos ortogonais, no sentido da maior variabilidade. Os componentes principais não devem ser correlacionados entre si e devem conter a maior parte da variância total explicada do conjunto de novas variáveis (FERREIRA, 2018).

Tendo um conjunto de p variáveis correlacionadas $X_1, X_2, \dots, X_p$, com covariância não nula entre a i -ésima e k -ésima variável definida por σ_{ik} , para $i \neq k = 1, 2, \dots$, e representando as p variáveis na forma vetorial por $X = [X_1, X_2, \dots, X_p]^T$, o vetor de médias $\mu = [\mu_1, \mu_2, \dots, \mu_p]^T$ e a matriz de covariâncias $\Sigma, p \times p$, o i -ésimo componente principal (scores) Y_i é obtido a partir de:

$$Y_i = e_i^T X$$

$$Y_i = e_{i1}X_1 + e_{i2}X_2 + \dots + e_{ip}X_p$$

Em que o vetor e_i corresponde à i -ésima combinação linear, para $i = 1, 2, \dots, p$.

O cálculo dos componentes principais é resumido na obtenção de autovalores y autovetores da matriz Σ , maximizando a forma quadrática da função $(\Sigma - \lambda_i I)e_i = 0$, em que I é uma matriz identidade $p \times p$, resultando na relação $(\Sigma e_i = \lambda_i e_i)$.

Segundo Ferreira (2018), a dispersão das observações em relação a qualquer um dos eixos coordenados depende da escala utilizada. Há casos em que o efeito da escala das variáveis faz com que haja predominância da representação dos componentes principais devido às grandes variâncias por parte de algumas variáveis. Nesses casos, é conveniente utilizar a matriz de correlação ρ ao invés da matriz de covariância Σ . Para isso, é necessário realizar uma padronização de posição e escala, a fim de obter a matriz de covariâncias do vetor aleatório Z , que será igual à matriz de correlações ρ . Utilizando essa padronização, o i -ésimo componente principal (scores) Y_i é obtido a partir de:

$$Y_i = e_i^T Z$$

$$Y_i = e_{i1}Z_1 + e_{i2}Z_2 + \cdots + e_{ip}Z_p$$

Modelando a matriz de correlação $\rho = P\Lambda P^T$, em que P é a matriz de autovetores ($p \times p$) e Λ é a diagonal de autovalores, se obtém a maximização da variância, resultando na equação $(\rho - \lambda_i I)e_i = 0$.

Os autovetores ou cargas e_i definem o sentido da rotação dos eixos, e os autovalores λ_i , definem a variação de cada componente principal. O primeiro componente principal (PC1) é o de maior variância, o PC2 é o que apresenta a segunda maior variância, e assim sucessivamente, até chegar ao componente principal de menor variância ($\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$). A variabilidade total das variáveis originais é igual à variabilidade total das componentes principais. Para obter $k < p$ componentes principais parcimoniosas é necessário considerar a proporção de explicação da variação por cada PC e a proporção de explicação da variância total (FERREIRA, 2018).

2.4.1 Número de componentes principais

Um critério para definir o número de PC é reter um número $k < p$ componentes principais que explicam pelo menos 70% da variação total. Em alguns casos mais rigorosos, é utilizado o critério de escolher um número $k < p$ componentes principais que explique pelo menos 80% da variação total (FERREIRA, 2018). A porcentagem da variação total explicada pelos $k < p$ primeiros componentes principais para os dados sem ou com padronização é dada por:

$$\rho_k^2 = \frac{\sum_{i=1}^k \lambda_i}{\sum_{i=1}^p \sigma_{ii}} \times 100 \text{ ou } \rho_k^2 = \frac{\sum_{i=1}^k \lambda_i}{p}$$

A porcentagem da explicação individual do k -ésimo componente principal para dados sem ou com padronização é:

$$p_k^2 = \frac{\lambda_k}{\sum_{i=1}^p \sigma_{ii}} \text{ ou } p_k^2 = \frac{\lambda_k}{p} \times 100$$

Outro critério para determinação do número de componentes a serem retidos é baseado na média aritmética dos autovalores da matriz de covariâncias ou matriz de correlação. Se a média aritmética dos autovalores da matriz de correlação for igual a 1, deve-se eliminar os componentes principais cujos autovalores sejam menores que 1, ou seja, menores que o valor médio de todos os autovalores não nulos.

$$\bar{\lambda} = \frac{\sum_{i=1}^p \lambda_i}{p} = \frac{p}{p} = 1$$

Outro procedimento utilizado é o gráfico de *scree plot*. Nesse gráfico, os autovalores (eixo y) são plotados com a ordem i (eixo x) das componentes principais. O número de componentes a serem retidas é o k componente situado no ponto no qual os autovalores tendem a se estabilizar (FERREIRA, 2018).

2.4.1.1 Número de variáveis

Jolliffe (2002) propôs alguns métodos para selecionar o número de k variáveis a partir da análise dos componentes principais. Dois desses métodos são baseados na associação das variáveis com os componentes principais de maior variância para selecionar variáveis. O método 1 considera o número de componentes com variância maior que a média das variâncias $\lambda_i > \bar{\lambda}$. Posteriormente, é selecionada uma única variável em cada PC, com base no maior coeficiente em módulo e que não tenha sido associada aos outros componentes. As demais $p-k$ variáveis devem ser eliminadas. No método 2, também são considerados os componentes com o critério do método 1. Logo, a cada um dos componentes principais de menor importância é

associada uma variável de maior coeficiente em módulo, que não tenha sido selecionada em outros componentes. Essas variáveis são eliminadas.

Após a obtenção dos autovetores segundo o número dos componentes principais, é possível conhecer a importância ou influência que cada variável original exerce sobre os componentes, calculando o grau de associação ou a correlação existente entre os componentes e as variáveis originais. O coeficiente de correlação entre Y_i e X_k é definido como:

$$\rho_{Y_i X_k} = \frac{\sqrt{\lambda_i} e_{ik}}{\sqrt{\sigma_{kk}}}$$

As correlações definem qual componente principal seria eficientemente utilizado para substituir as variáveis originais. E as variáveis mais relacionadas com os componentes podem ser utilizadas para nomeá-los ou dar um significado prático na análise (FERREIRA, 2018).

2.4.2 Algoritmos de *Clustering*

A delimitação de ZM em áreas agrícolas é uma técnica economicamente viável em AP. Os algoritmos de *clustering* são métodos de agrupamento de objetos semelhantes no mesmo grupo homogêneo, sendo dissimilares entre grupos diferentes. Para definir as ZM, os métodos mais utilizados são os algoritmos de análise cluster K-means e fuzzy K-means (GAVIOLI et al., 2019; LIKAS; VLASSIS; VERBEEK, 2020).

2.4.2.1 K-means

Seja $X = \{x_i\}, i = 1, \dots, n$ um conjunto de pontos a serem agrupados em um conjunto de k clusters $C = \{C_k, k = 1, \dots, K\}$, o algoritmo K-means gera uma partição que minimiza o erro quadrático entre a média empírica de um cluster (centróide ou centro de um grupo) e os pontos dentro do cluster. Seja μ_k a média do cluster C_k , o erro quadrado entre μ_k e os pontos no cluster C_k são definidos como: $J(C_k) = \sum_{x_i \in C_k} \|x_i - \mu_k\|^2$. O algoritmo K-means tem por objetivo minimizar a expressão da soma do erro quadrático em todos os grupos k . Neste procedimento, a distância euclidiana é usada como norma para expressar a distância entre as observações (JAIN, 2010).

$$J(C) = \sum_{k=1}^K \sum_{x_i \in C_k} \|x_i - \mu_k\|^2$$

As etapas gerais do algoritmo K-means são:

1. Selecionar uma partição inicial com k cluster;
2. Gerar uma nova partição, atribuindo cada padrão ao centro de cluster mais próximo, ou seja, atribui-se cada objeto ao grupo cujo centróide possua maior similaridade com o objeto;
3. Calcular novos centros de cluster de cada grupo;
4. Repetir as etapas 2 e 3 até que a associação do cluster se estabilize.

2.4.2.2 Algoritmo Fuzzy K-means

A classificação de ZM pelo método fuzzy K-means tem sido utilizada para estudos de rendimentos e atributos do solo de culturas, tais como milho, soja, algodão, pessegueiro e tabaco (GAVIOLI et al., 2016; GILI et al., 2017; LI et al., 2007; OLDONI et al., 2019; XIN-ZHONG et al., 2009). Além disso, essa classificação também tem sido utilizada em estudos de CE e atributos do solo de cultivos de soja, milho e arroz (PERALTA; COSTA, 2013; TRIPATHI et al., 2015). Outros estudos incluem a variável de índice de solo e vegetação por diferença normalizada NDVI, pela explicação da variabilidade espacial do rendimento em campos de produção de algodão e uva (LI et al., 2007; OHANA-LEVI et al., 2019).

O fuzzy K-means é uma extensão do método K-means, em que cada ponto do conjunto de dados pode fazer parte de vários clusters com um valor de associação, chamado atribuição suave (JAIN, 2010). Esse algoritmo usa um expoente de ponderação para controlar o grau em que o compartimento de associação ocorre entre os clusters (FRIDGEN et al., 2004).

Para que uma observação individual pertença a um cluster específico, é necessário ter uma medida de similaridade, que é um procedimento usado para determinar a semelhança de uma observação com determinado centro de cluster (centróide). A similaridade aumenta à medida em que a diferença entre uma observação e a média do cluster diminui (LIKAS; VLASSIS; VERBEEK, 2020).

Uma das medidas comumente usadas é a distância euclidiana, que fornece peso igual a todas as variáveis medidas e é caracterizada por ser sensível a variáveis correlacionadas (FRIDGEN et al., 2004). No processo de agrupamento, existem três matrizes:

- Matriz de dados Y , contendo n observações com p variáveis de classificação;
- Matriz centroide do cluster V , contendo c centroides delimitados no espaço de recurso definido pelas variáveis de classificação p ;
- Matriz de associação difusa U , contendo valores de associação para cada cluster em V para cada observação em Y , delimitada pelas restrições de todo $i = 1$ a c e todo $k = 1$ a n que: $\mu_{ik} \in [0,1]$, $1 \leq i \leq c$, $1 \leq k \leq n$ e $\sum_{i=1}^c \mu_{ik} = 1$, $1 \leq k \leq n$.

O algoritmo permite localizar soluções mínimas para uma função objetivo selecionada. Nas ZM, a função objetivo ponderada é comumente usada dentro dos grupos de erros quadráticos (LI et al., 2007).

$$J_m(U, v) = \sum_{k=1}^n \sum_{i=1}^c (\mu_{ik})^m (d_{ik})^2$$

Em que m é o expoente de imprecisão ($1 \leq m < \infty$) e $(d_{ik})^2$ é a distância ao quadrado no espaço de características entre y_k e v_i .

O expoente m controla a quantidade de compartilhamento de associação entre os clusters. À medida em que m tende ao infinito, os agrupamentos se tornam menos distintos devido à quantidade de compartilhamento entre grupos.

A distância no espaço entre uma observação em Y e um centróide de cluster em V é expressa por: $(d_{ik})^2 = \|y_k - v_i\|^2 = (y_k - v_i)' \mathbf{A} (y_k - v_i)$, onde y_k é a observação de dados k , constituída por p variáveis de classificação; v_i é o centróide de cluster i , constituído por variáveis de classificação p ; e $\mathbf{A}$ é a matriz de peso positiva definida, de norma induzida, de tamanho $p \times p$, que define o procedimento de normalização para a distância euclidiana.

2.4.3 Índices de validação de ZM

Diversos autores que utilizaram a análise fuzzy K-means validaram as partições de ZM com os índices de partição difuso (FPC) e entropia de partição normalizada (NCE). Tais índices são os mais citados na literatura como indicadores do número ideal de clusters ou número de classes de agrupamentos para criar ZM.

O índice FPC calcula o grau de imprecisão entre os grupos fuzzy gerados a partir de um conjunto de dados. Quanto mais próximo de 1, mais distintos são os grupos, com menor grau de compartilhamento de elementos entre os grupos gerados. Por sua vez, o índice NCE fornece

a estimativa da quantidade de desorganização de grupos. Valores de NCE próximos a 1 indicam maior desorganização de grupos (TRIPATHI et al., 2015). O número de clusters é definido quando o FPC está no valor máximo e NCE está no valor mínimo, ou seja, menor compartimento de associação de elementos entre grupos e maior organização de grupos. (LI et al., 2007).

O coeficiente de partição difuso (FPC) é um índice que mede a clareza da partição. Os valores máximos do índice indicam uma boa partição, ou seja, um melhor agrupamento do conjunto de dados (MEYER et al., 2021). O FPC mostra com que clareza um conjunto de dados é descrito por um modelo de agrupamento, o que significa que valores próximos a 1 têm maior probabilidade de pertencer a determinado cluster, e não aos demais. Portanto, quanto mais altos os valores de FPC, melhor será o agrupamento (HE, 2018).

$$FPC = \frac{1}{n} \sum_{k=1}^n \sum_{i=1}^c u_{ik}^2$$

$$NCE = \frac{-\sum_{k=1}^n \sum_{i=1}^c u_{ik} \log(u_{ik}) / n}{\log c}$$

Em que c é o número de grupos; n é o número de elementos no conjunto de dados; e u_{ij} é o valor da distância entre observações em um centróide de cluster.

2.4.4 Análise de variância (ANOVA)

Para avaliar a heterogeneidade entre ZM, a análise de variância pode ser utilizada. Li et al. (2007) e Xin-Zhong et al. (2009) utilizaram a ANOVA para verificar a diferença estatística de clusters definidos pelos índices FPC e NCE. Nesse trabalho, a definição de ZM foi realizada com o algoritmo Fuzzy K-means. Gavioli et al. (2016), por sua vez, identificaram uma recomendação de classes ou clusters combinada com os índices FPC, NCE e ANOVA. Usando a ANOVA, Tripathi et al. (2015) e Xin-Zhong et al. (2009) avaliaram a efetividade da combinação de ACP e o algoritmo Fuzzy K-means para delinear as ZM. Quanto maiores as diferenças, mais a validade da divisão foi confirmada. Em outro estudo, Peralta e Costa (2013) mostraram que a variável CE delimita com sucesso as ZM associadas com distribuição espacial das propriedades de solo e alguns nutrientes. Detectada alguma diferença estatisticamente significativa entre clusters, os valores médios na análise pós-ANOVA de Tukey foram

utilizados para identificar o comportamento da variável resposta de interesse em relação às ZM (GAVIOLI et al., 2019).

3 MATERIAL E MÉTODOS

Nesta seção do documento, foram descritas as bases de dados da área de estudo e os métodos utilizados para a análise estatística.

3.1 Descrição da área

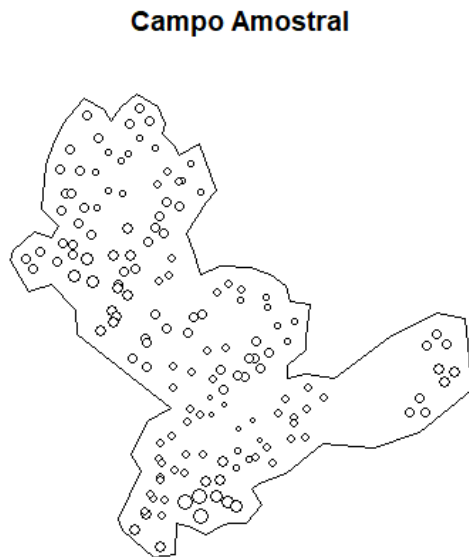
Os dados utilizados neste estudo correspondem a uma produção de cana-de-açúcar em uma área de 1516,33 ha, localizada na fazenda Tehuantepec no município de Santa Lucía Cotzumalguapa, Escuintla, Guatemala (Figura 2). Na safra de 2019-2020, foram amostrados 153 pontos georreferenciados (Figura 3) para 17 atributos físico-químicos do solo: potencial hidrogeniônico, fósforo, cobre, zinco, ferro, manganês, boro, matéria orgânica, capacidade de troca catiônica, cálcio, magnésio, sódio, potássio, saturação de bases, argila, silte e areia.

Figura 2 - Mapa de localização do município Santa Lucía Cotzumalguapa, Escuintla, Guatemala.



FONTE: Google Earth (2021).

Figura 3 - Campo amostral dos 153 pontos de amostragem do solo georreferenciados na fazenda produtora de cana-de-açúcar Tehuantepec, localizada na região sul da Guatemala, no município de Santa Lucía Cotzumalguapa, Escuintla.



3.2 Estatística descritiva

Todos os atributos do solo foram submetidos à análise exploratória descritiva, obtendo-se gráficos e os valores de média aritmética, mediana, mínimo, máximo, desvio padrão, coeficiente de variação, assimetria e curtose. Foi realizada a análise de correlação pelo método de Pearson, a fim de determinar a relação linear existente entre as variáveis em estudo (TRIPATHI et al., 2015).

Para delimitar as ZM, foram utilizadas as seguintes metodologias: análise Geoestatística, análise de componentes principais (ACP) e análise de agrupamento por meio do algoritmo Fuzzy K-means.

3.3 Análise Geoestatística

A determinação da dependência espacial de cada variável foi analisada pela análise estrutural ou análise variográfica, considerando para seu ajuste as variáveis que não apresentaram efeito pepita puro e com comportamento característico da teoria variográfica. Também foi usada a classificação de dependência espacial segundo Cambardella et al. (1991), determinada pela relação em porcentagem do efeito pepita em relação ao patamar, sendo forte para valores menores até 25%, moderada entre 25% e 75% e fraca para valores maiores que 75%.

Os modelos esférico e exponencial foram ajustados aos variogramas experimentais utilizando o método dos mínimos quadrados. Para avaliar a qualidade do ajuste do modelo teórico do variograma, foi realizada a validação cruzada. Foram considerados a média e o desvio padrão dos erros padronizados, o intercepto e o coeficiente angular da reta da regressão, os quais devem ser próximos a 0, 1, 0 e 1, respectivamente.

Os modelos ajustados foram utilizados na etapa da interpolação por krigagem ordinária para estimar os valores desses atributos do solo em locais não amostrados, bem como para a geração de mapas de distribuição espacial.

3.4 Análise de componentes principais

A análise de componentes principais foi aplicada aos valores preditos dos atributos do solo, obtidos pela técnica de interpolação por krigagem. Como as variáveis têm diferentes unidades e escala de medição, foi utilizada a matriz de correlação ao invés da matriz de covariâncias para a redução de dimensionalidade. Foram obtidos os autovetores e autovalores para determinar o número de componentes principais, os CPs foram selecionados com base na explicação acumulada maior que 70% da variância total e de cada autovalor dos PCs maiores que 1. Os escores dos PCs e as variáveis selecionadas foram obtidos e plotados para gerar o mapa de contorno de cada componente.

3.5 Análise de agrupamento para delinear zonas de manejo

A análise cluster é um método de agrupamento de objetos semelhantes no mesmo grupo homogêneo, sendo dissimilar entre grupos diferentes. O algoritmo cluster Fuzzy K-means foi utilizado para gerar zonas de manejo. Esse algoritmo usa um expoente de ponderação para controlar o grau em que o compartimento de associação ocorre entre os clusters (FRIDGEN et al., 2004). A distância euclidiana foi utilizada para calcular a distância dos pontos das observações dos dados em relação aos pontos centrais dos cluster. Para a rotina da análise, foi utilizada a metodologia de Córdoba (2014), considerando o mínimo de 2 e máximo de 6 zonas de manejo, 100 interações e expoente difuso de 1,3.

O número de clusters foi definido com os critérios do coeficiente de partição difusa (FPC) e entropia de partição normalizada (NCE). O coeficiente de partição é um índice que mede a clareza da partição e não considera o próprio conjunto de dados. Os valores máximos do índice indicam uma boa partição, ou seja, melhor eficiência de agrupamento do conjunto de

dados (MEYER et al., 2021). O FPC mostra com que clareza um conjunto de dados é descrito por um modelo de agrupamento, o que significa que valores próximos a 1 possuem maior probabilidade de pertencer a um determinado cluster, e não aos demais. Portanto, quanto mais altos os valores de FPC, melhor será o agrupamento (HE, 2018).

O índice NCE, por sua vez, fornece a estimativa da quantidade de desorganização de grupos. Valores próximos a 1 indicam maior desorganização de grupos levando em consideração os próprios dados, enquanto os valores mínimos indicam uma boa partição (MEYER et al., 2021; TRIPATHI et al., 2015). Para tornar a interpretação do FPC semelhante ao NCE, calculou-se o inverso dos valores do índice FPC. Assim, o melhor número de classes ZM é definido quando ambos os índices assumem valores mínimos (LI et al., 2007; CÓRDOBA, 2014).

Todas as análises estatísticas foram realizadas no *software* R e RStudio (R CORE TEAM, 2020). Os pacotes utilizados foram *geoR* e *ppclust* (CEBECI et al., 2018; RIBEIRO JUNIOR et al., 2020).

4 RESULTADOS E DISCUSSÃO

4.1 Estatística descritiva

As estatísticas descritivas dos atributos do solo estão resumidas na Tabela 2. Foi encontrado que o solo da área de estudo é ligeiramente ácido, com pH variando de 5,4 a 6,80. A estrutura do solo estudado é composta em termos de valores médios por 53,82% de areia, 32,77% de silte e 13,41% de argila. A média do conteúdo de matéria orgânica encontrada foi de 4,85%, indicando um conteúdo moderado (CENGICANÑA, 2012). As variáveis do solo P, Cu, Zn, Fe, Mg e K apresentaram grande variabilidade nos dados, com coeficientes de variação acima de 50%. Para Mn, B, MO, CIC, Ca, Na, SB, Argila, Silte e areia, os CV variaram de 19,81% a 48,38%. Por sua vez, o pH teve a menor variação, com 4,19%.

O grau de assimetria de uma distribuição de dados é caracterizado pelo coeficiente de assimetria, enquanto o grau de achatamento é quantificado pelo coeficiente de curtose. Uma distribuição normal tem valores desses dois coeficientes iguais a zero. No nosso estudo, os coeficientes de assimetria variaram de -0,73 até 4,41, as distribuições dos dados de Cu, Zn, Fe, Ca, Mg, Na, K e argila tiveram assimetria positiva, com coeficientes maiores a 1, indicando que foram encontrados valores altos e muito altos desses elementos em alguns pontos. Os coeficientes de curtose oscilaram de -1,43 à 23,00, os atributos pH, Cu, Zn, Fe, Ca, Na e K

apresentaram coeficientes de curtose maiores que 1,00, destacando-se o Cu (23,00) e K (12,47), enquanto o P foi o único atributo que apresentou coeficiente de curtose negativo, menor que -1,00. As distribuições não apresentaram normalidade pelo teste de Shapiro-Wilk a um nível de 0,05 de significância. No entanto, a pressuposição de normalidade não é necessária para as análises a seguir. Os gráficos das distribuições dos dados podem ser observados nas Figuras 18, 19 e 20 do apêndice.

Tabela 2 - Estatística descritiva para os atributos físicos e químicos do solo da área produtora de cana-de açúcar com medições georreferenciadas em 153 locais amostrados.

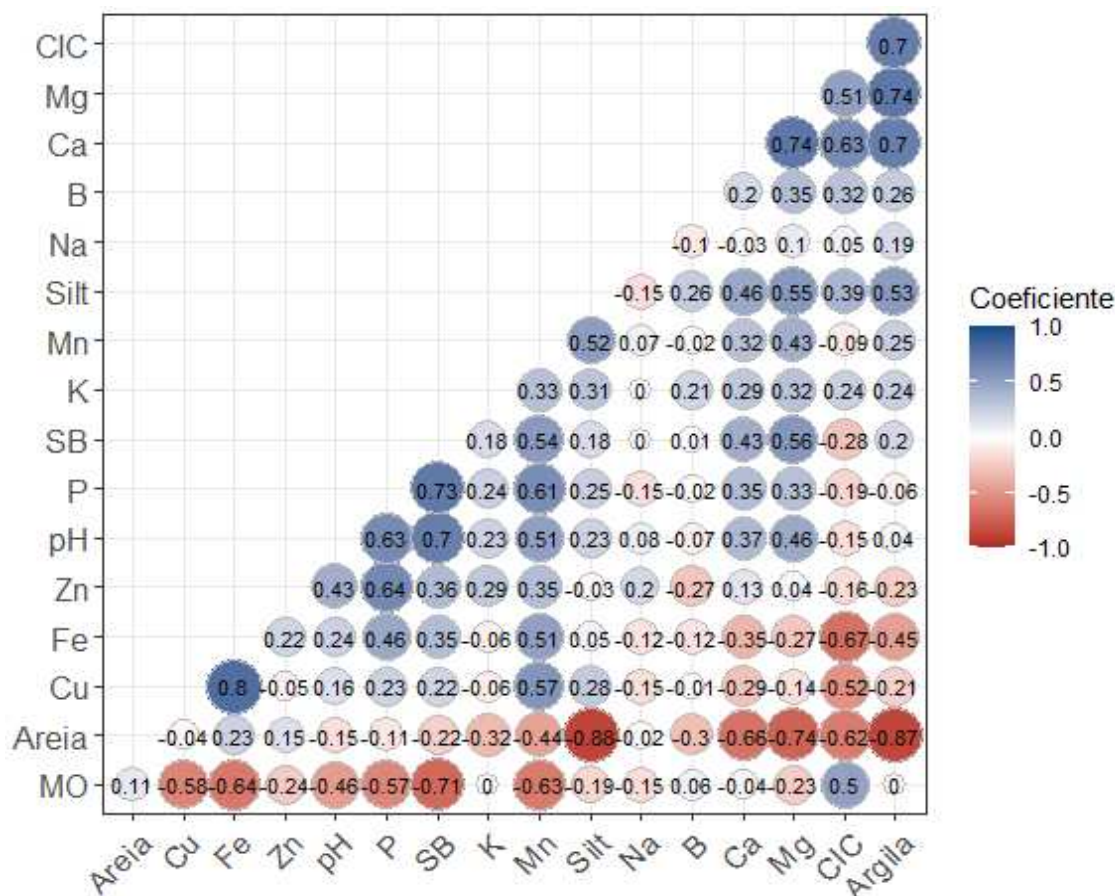
Atributo	Média aritmética	Mediana	Desvio padrão	CV %	Min	Máx	As	Curtose	p-valor*
pH	5,97	6,00	0,25	4,19	5,40	6,80	0,49	1,23	<0,001
P (ppm)	80,42	76,64	57,55	71,56	0,98	166,84	0,01	-1,43	<0,001
Cu (ppm)	0,61	0,50	0,91	147,86	0,10	6,00	4,41	23,00	<0,001
Zn (ppm)	6,70	6,00	3,73	55,63	2,00	21,50	1,48	3,31	<0,001
Fe (ppm)	11,59	7,00	11,94	103,06	0,50	60,00	2,09	4,68	<0,001
Mn (ppm)	24,08	24,00	11,65	48,38	6,50	63,00	0,77	0,93	<0,001
B (ppm)	0,85	0,86	0,21	24,37	0,48	1,30	0,03	-0,72	0,004
MO (%)	4,85	4,93	1,57	32,33	1,85	7,99	-0,09	-0,59	0,008
CIC (meq/100g)	29,61	30,38	8,31	28,08	7,31	47,92	-0,73	0,63	<0,001
Ca (meq/100g)	11,37	11,23	4,29	37,70	4,24	26,20	1,01	1,73	<0,001
Mg (meq/100g)	2,99	2,43	1,63	54,71	0,66	7,61	1,06	0,47	<0,001
Na (meq/100g)	0,30	0,30	0,12	41,06	0,14	0,70	1,55	3,09	<0,001
K (meq/100g)	0,52	0,33	0,47	90,34	0,09	2,95	3,04	12,47	<0,001
SB (% ⁻)	52,90	55,85	14,65	27,69	24,59	90,75	0,20	-0,29	0,002
Argila (%)	13,41	11,42	6,45	48,09	4,37	32,42	1,02	0,44	<0,001
Silte (%)	32,77	31,50	6,49	19,81	20,24	47,54	0,42	-0,28	<0,001
Areia (%)	53,82	57,83	11,32	21,03	30,54	74,64	-0,34	-0,82	<0,001

Abreviações. CV %: coeficiente de variação em porcentagem, Mín: mínimo, Máx: máximo, As: assimetria, pH: potencial hidrogeniônico, P: fósforo, Cu: cobre, Zn: zinco, Fe: ferro, Mn: manganês, B: boro, MO: matéria orgânica, CIC: capacidade de troca catiônica, Ca: cálcio, Mg: magnésio, Na: sódio, K: potássio, SB: saturação de bases. *O p-valor corresponde ao teste de normalidade de Shapiro-Wilk para cada atributo do solo.

4.2 Análise de correlação

Foi obtida a matriz de correlação linear de Pearson para os atributos do solo (Figura 4). Segundo Cohen (1988), coeficientes de correlação para valores absolutos entre 0,10 e 0,29 podem ser considerados baixos; valores entre 0,30 e 0,49 podem ser considerados como médios; e valores entre 0,50 e 1 podem ser interpretados como altos. A produção das culturas é limitada por fatores agronômicos como ervas daninhas, pragas, doenças e déficit de nutrientes. Assim, o rendimento potencial é alcançado quando o cultivo tem condições ótimas de água e nutrientes na ausência de fatores limitantes (CENGICAÑA, 2012).

Figura 4 - Matriz de correlação linear de Pearson para os atributos do solo da área de estudo de produção de cana-de-açúcar.



A capacidade de reter água no solo depende da classe da textura presente. Na área de estudo houve predominância do conteúdo de areia (53,82%). Os conteúdos de silte e argila foram 32,77% e 13,41% respectivamente, o que indica que o solo é, de forma geral, franco-arenoso. Solos com alto conteúdo de areia apresentam baixa retenção de água e, portanto, perdas de nutrientes devido à lixiviação (CENGICANA, 2012). Na matriz de correlação, o atributo físico de areia obteve a maioria dos coeficientes negativos, de moderada a alta correlação, indicando que o aumento do conteúdo de areia pode estar associado à redução de silte, argila, Mg, Ca, CIC, Mn, K, B.

No caso do pH, houve uma relação direta com a SB, refletido no coeficiente de correlação de 0,70. As bases Ca^+ , 9Mg^+ e K^+ apresentaram correlação positiva com SB e a base Na^+ obteve correlação nula. O elemento Na apresentou baixas correlações com os demais atributos. A maioria dos coeficientes de correlação dos atributos com outras variáveis foram estatisticamente significativos (Tabela 11, apêndice).

4.3 Análise Geoestatística

Foram calculadas as semivariâncias, construídos os variogramas e identificados os melhores modelos para seu ajuste. Visualmente, nas Figuras 5, 6 e 7 todas as variáveis apresentaram dependência espacial, com exceção do atributo potássio (K), que teve comportamento próximo ao efeito pepita puro. Na Tabela 3, estão apresentados os parâmetros da análise variográfica e os resultados da validação cruzada. Com base a classificação de dependência espacial de Cambardella et al. (1994), 13 das 17 variáveis apresentaram forte dependência espacial. Os atributos fósforo (P), boro (B) e argila apresentaram moderada dependência no espaço e fraca dependência espacial para potássio (K). O atributo potássio (K) foi o que apresentou menor coeficiente de determinação (0,16), valor que pode ser explicado pela fraca dependência espacial.

Figura 5 - Variograma empírico e variograma teórico ajustado pelo modelo esférico para os atributos potencial hidrogeniônico, fósforo, cobre, zinco, ferro e manganês.

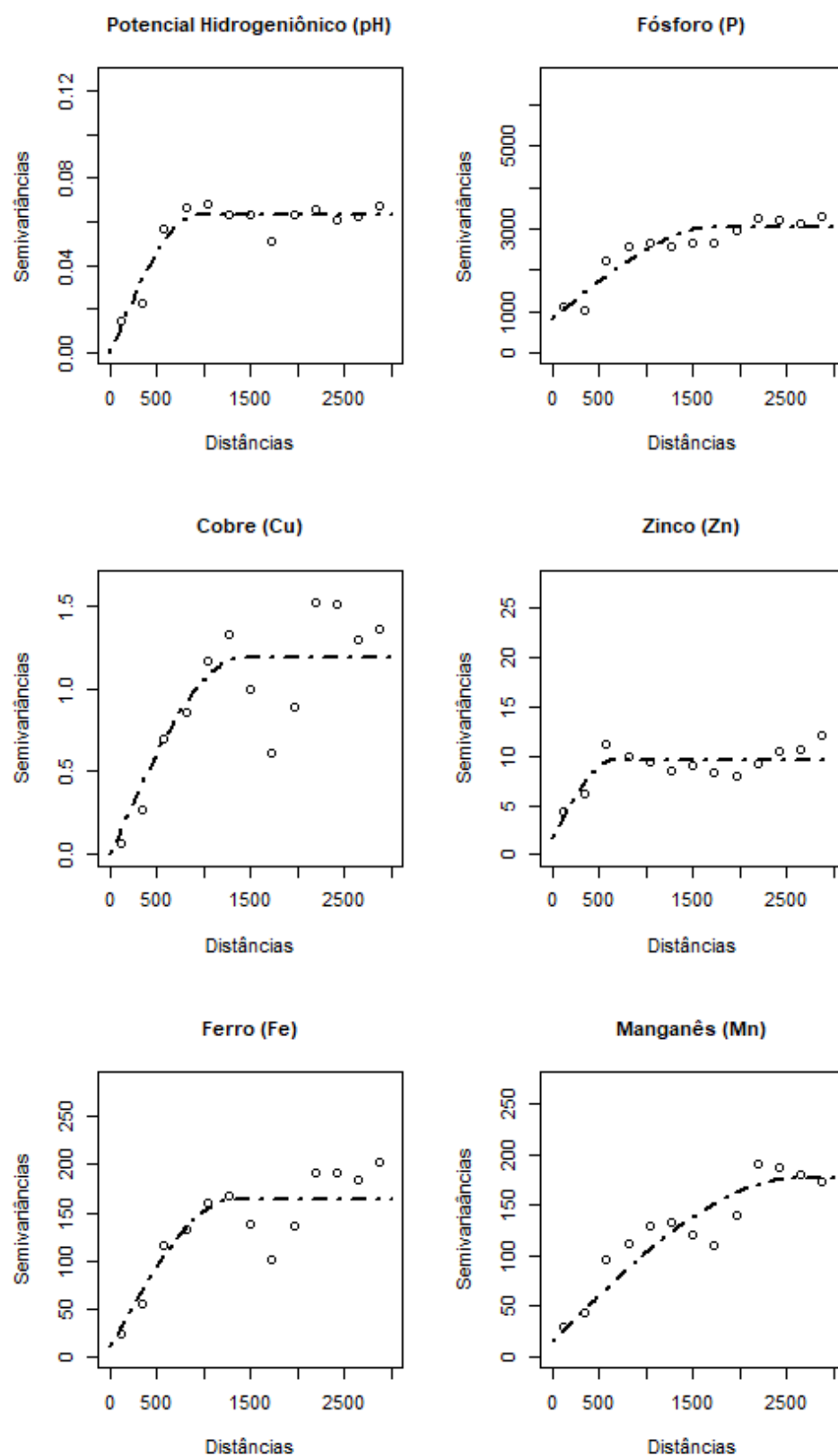


Figura 6 - Variograma empírico e variograma teórico ajustado pelo modelo esférico para os atributos boro, matéria orgânica, capacidade de troca catiônica, cálcio, magnésio e modelo exponencial para a variável sódio.

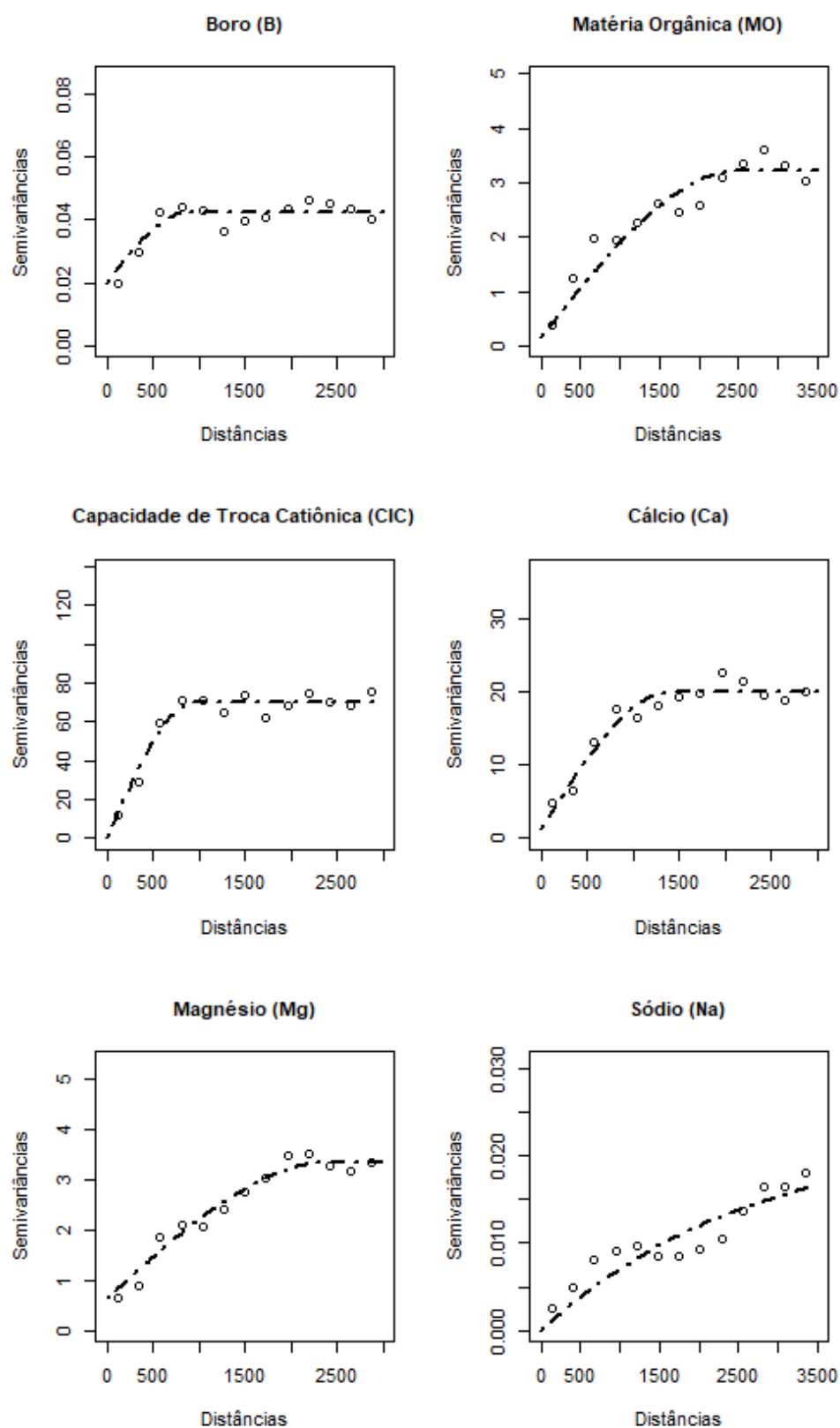


Figura 7 - Variograma empírico e variograma teórico ajustado pelo modelo esférico para os atributos potássio, saturação de bases, argila, silte e areia.

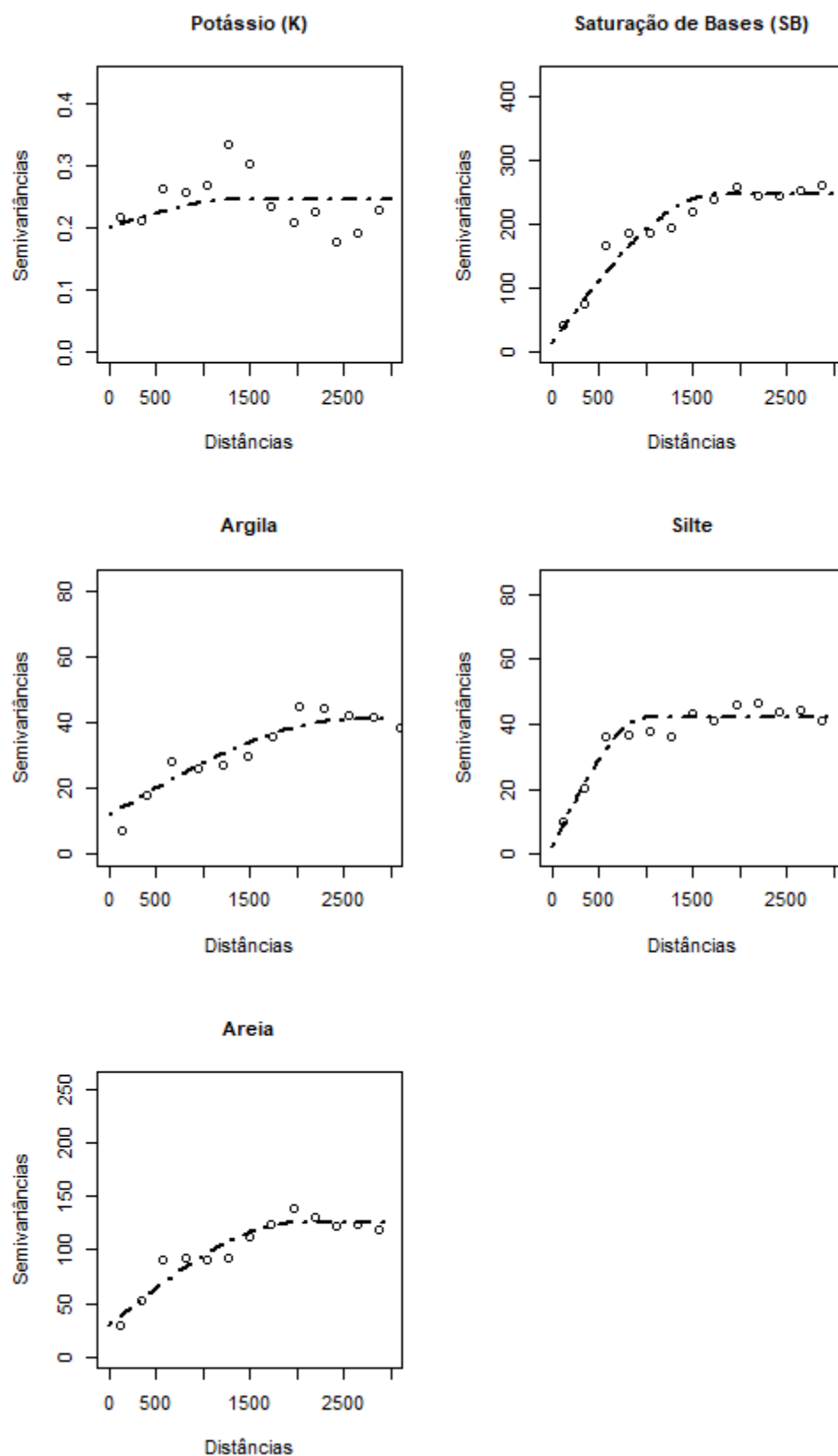


Tabela 3 - Parâmetros variográficos e resultados da validação cruzada dos valores preditos pela técnica de interpolação Krigagem e valores observados dos atributos do solo.

Atributo	Modelo	Parâmetros variográficos				Validação cruzada				
		a (m)	C_0	$C_0 + C_1$	RDE%	MSE	RSE	β_0	β_1	R^2
pH	Esférico	1.099	0,00	0,07	0,00	0,00	0,97	-0,29	1,05	0,71
P	Esférico	1.815	828,00	3.040,00	27,24	0,00	0,89	-7,50	1,10	0,65
Cu	Esférico	1.557	0,00	1,10	0,00	0,00	0,85	-0,04	1,05	0,76
Zn	Esférico	889	1,70	10,00	17,00	0,00	0,93	-2,38	1,36	0,64
Fe	Esférico	1.312	9,70	155,00	6,26	0,00	0,91	-0,74	1,06	0,73
Mn	Esférico	1.811	15,00	152,00	9,87	0,00	1,03	-1,71	1,07	0,69
B	Esférico	1.070	0,02	0,04	50,00	0,00	0,89	-0,39	1,46	0,43
MO	Esférico	1.170	0,17	3,50	4,86	0,01	0,94	-0,48	1,01	0,72
CIC	Esférico	1.097	0,00	71,00	0,00	0,01	0,90	-2,78	1,09	0,73
Ca	Esférico	1.401	1,30	20,00	6,50	0,00	0,94	-0,57	1,05	0,72
Mg	Esférico	2.555	0,64	3,30	19,39	0,01	0,84	-0,29	1,01	0,70
Na	Exponencial	1.603	0,00	0,02	0,00	0,00	1,31	0,01	1,03	0,76
K	Esférico	1.368	0,20	0,26	76,92	0,00	0,91	-0,33	1,64	0,16
SB	Esférico	1.505	15,00	233,00	6,44	0,01	0,99	-2,01	1,04	0,73
Argila	Esférico	3.016	12,00	43,00	27,91	0,01	1,03	-2,51	1,19	0,64
Silte	Esférico	1.005	2,40	40,00	6,00	0,00	0,65	-1,92	1,06	0,59
Areia	Esférico	1.945	30,00	125,00	24,00	0,00	0,97	-8,37	1,16	0,62

pH: potencial hidrogeniônico, P: fósforo, Cu: cobre, Zn: zinco, Fe: ferro, Mn: manganês, B: boro, MO: matéria orgânica, CIC: capacidade de troca catiônica, Ca: cálcio, Mg: magnésio, Na: sódio, K: potássio, SB: saturação de bases, a (m): alcance em metros, C_0 : efeito pepita, $C_0 + C_1$: patamar, RDE %: razão de dependência espacial em %, MSE: erro quadrático médio, RSE: raiz do erro padrão, β_0 : intercepto, β_1 : coeficiente angular, R^2 : coeficiente de determinação.

O modelo esférico foi o mais adequado para o ajuste de todas as variáveis, exceto para o sódio (Na), para o qual o modelo exponencial se ajustou melhor, com um patamar assintótico de 0,02. Diversos autores relatam que o modelo esférico se ajusta melhor aos dados do variograma experimental na maioria das propriedades do solo (GILI et al., 2017; OLDONI et al., 2019; TRIPATHI et al., 2015; XIN-ZHONG et al., 2009). O alcance ou distâncias de dependência espacial variou de 889 m (variável zinco) a 3.016 m (variável argila).

Para avaliar a qualidade do ajuste do modelo teórico, foi feita a validação cruzada. A média padronizada dos erros para todas as variáveis foram próximas a 0 e o desvio padrão padronizado foi próximo a 1 para 15 dos 17 atributos. Os valores mais afastados foram do elemento sódio (Na), com 1,31 e silte, com 0,65. O intercepto, que segundo Santos (2010) preferencialmente deve ser zero, mostrou ter valores baixos (-8,37 para areia e -7,50 para fósforo), indicando subestimação da modelagem. Por sua vez, o coeficiente angular da regressão entre valores preditos e observados das propriedades físico-químicas foi próximo ao valor ideal (1). Os valores de coeficiente angular que mais se afastaram do ideal foram obtidos para os atributos potássio (K), boro (B) e zinco (Zn), com 1,64, 1,46 e 1,36, respectivamente.

O atributo potássio (K) não apresentou dependência espacial e, portanto, foram consideradas 16 das 17 variáveis físico-químicas do solo para fazer a interpolação por krigagem. A grid utilizada foi de 300 x 300, gerando um total de 90.000 valores interpolados. Os mapas gerados pelo método de krigagem estão apresentados nas Figuras 8, 9 e 10. Valores baixos predominaram nos mapas de potencial hidrogeniônico (pH), cobre (Cu), zinco (Zn), ferro (Fe), manganês (Mn), cálcio (Ca), magnésio (Mg), sódio (Na) e argila. Diferentes padrões de distribuição espacial foram observados nos mapas de fósforo (P), boro (B), matéria orgânica (MO), capacidade de troca catiônica (CIC), saturação de bases (SB), silte e areia.

Figura 8 - Mapas de distribuição espacial obtidos a partir da krigagem para os atributos: potencial hidrogeniônico, fósforo, cobre, zinco, ferro e manganês.

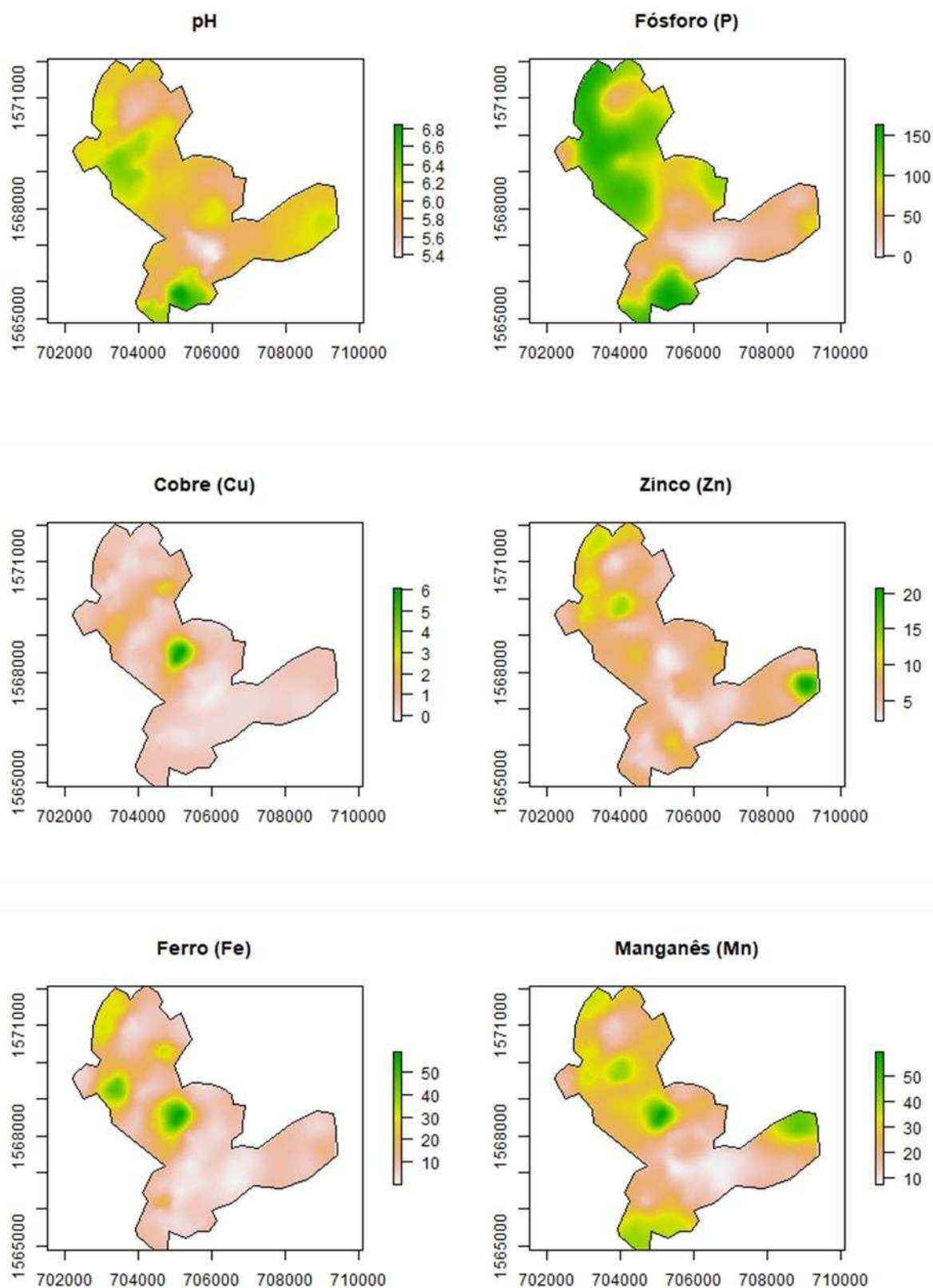


Figura 9 - Mapas de distribuição espacial obtidos a partir da krigagem para os atributos: boro, matéria orgânica, capacidade de troca catiônica, cálcio, magnésio e sódio.

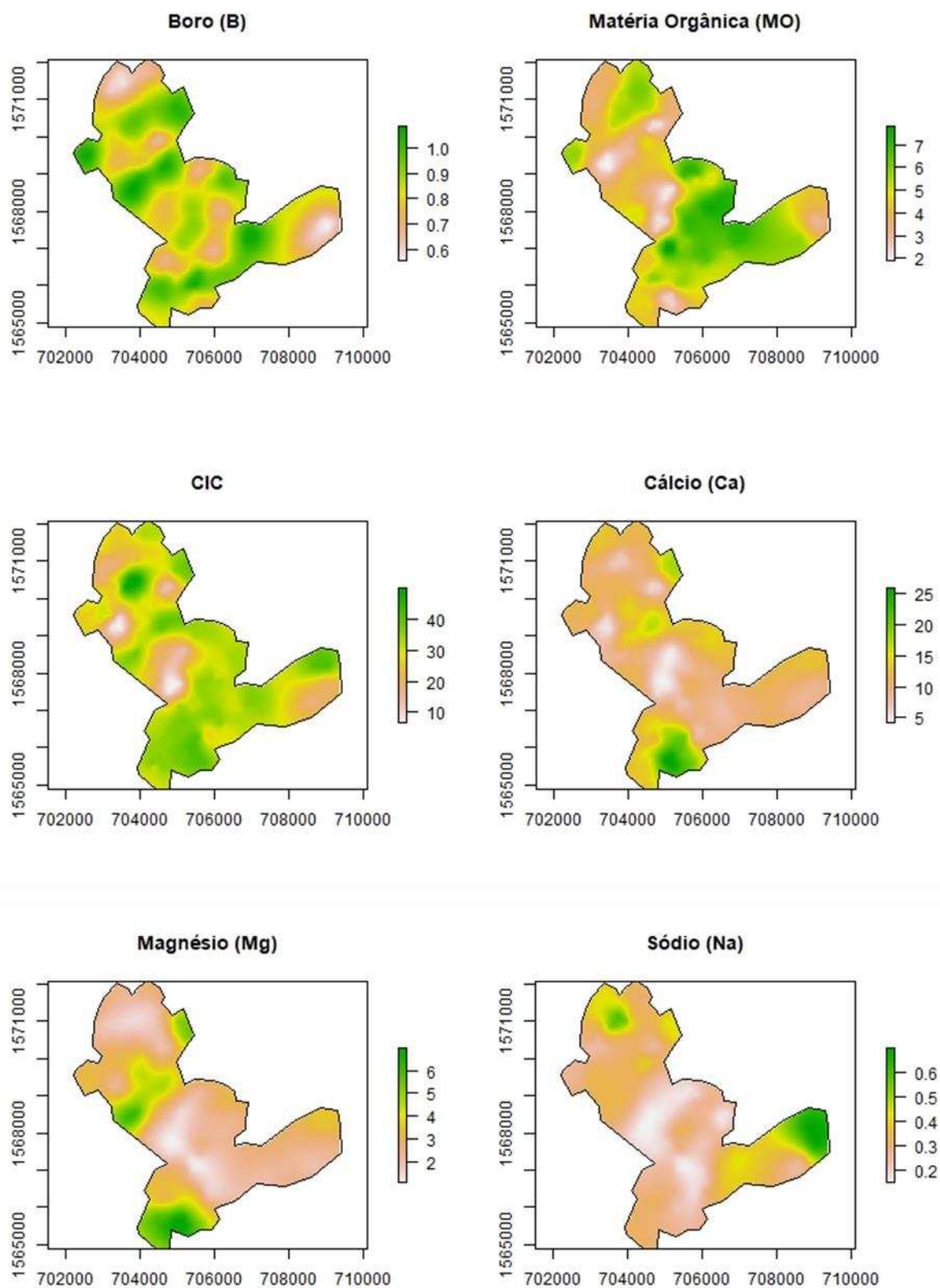
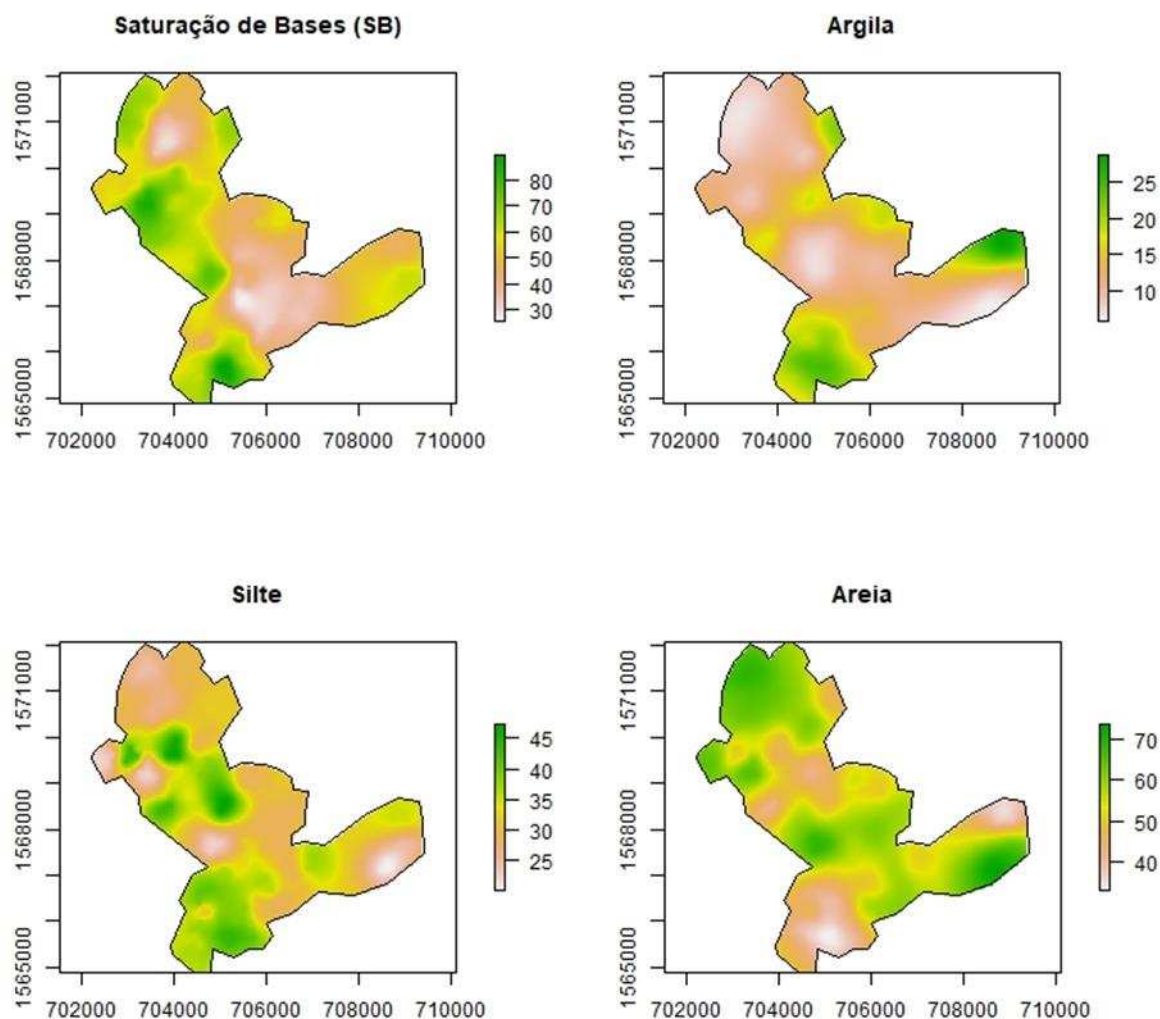


Figura 10 - Mapas de distribuição espacial obtidos a partir da krigagem dos atributos: potássio, saturação de bases, argila e areia.

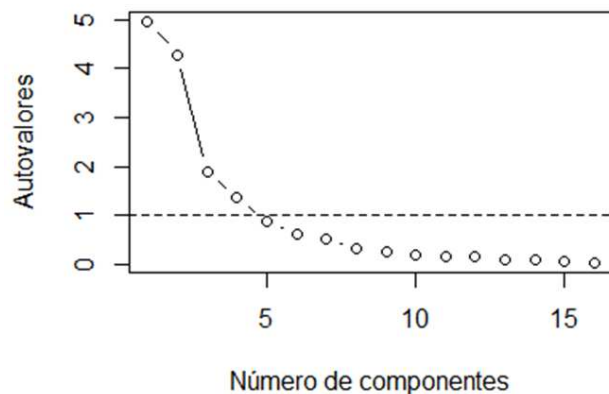


4.4 Análise de componentes principais (ACP)

Para reduzir a dimensionalidade dos dados de atributos físico-químicos do solo, foi aplicada a análise de componentes principais. A ACP visa, a partir de variáveis originais relacionadas, obter variáveis latentes identificando uma combinação linear ortogonal. O primeiro componente principal (PC1) foi o de maior variância, o segundo (PC2) apresentou a segunda maior variância, e assim sucessivamente. Com a ACP, foram obtidos os autovalores, a proporção e porcentagem da variabilidade explicada por cada componente e cumulativa. De acordo com o critério de Ferreira (2018) e analisando o gráfico de *scree plot* (Figura 11), observou-se que 4 componentes principais poderiam ser retidos, já que os autovalores estabilizaram nesse ponto. A partir do componente 5, os autovalores foram menores que um e

as porcentagens da variância não mudaram muito entre os componentes principais seguintes até chegar ao último, com menor variância.

Figura 11 - Gráfico de *Scree plot* dos autovalores em função do número de componentes principais das 16 variáveis do solo que apresentaram dependência espacial.



Considerando autovalores maiores que um e a variação acumulada superior a 70%, foram escolhidos os quatro primeiros componentes (PC1, PC2, PC3 e PC4), que explicaram 78,30% dos dados (Tabela 4).

Tabela 4 - Resultados da análise de componentes principais aplicada aos dados interpolados pela técnica Krigagem dos atributos do solo com variabilidade espacial.

Componente Principal	Autovalores	Variação (%)	Variação acumulada (%)
PC1	4,95	30,97	30,97
PC2	4,29	26,82	57,79
PC3	1,90	11,89	69,68
PC4	1,38	8,63	78,30
PC5	0,89	5,58	83,88
PC6	0,61	3,80	87,68
PC7	0,52	3,23	90,91
PC8	0,34	2,14	93,06
PC9	0,28	1,73	94,79
PC10	0,21	1,31	96,09
PC11	0,16	1,02	97,11
PC12	0,15	0,95	98,07
PC13	0,11	0,68	98,75
PC14	0,09	0,54	99,29
PC15	0,06	0,41	99,69
PC16	0,05	0,31	100,00

De acordo com a metodologia de eliminar variáveis, proposta por Jolliffe (2002), o PC4 foi considerado de menor importância na primeira análise de componentes principais por apresentar autovalor maior que um (1,38). Sendo assim, a variável sódio foi eliminada, uma vez que foi o atributo com maior valor dos autovetores, em módulo, associada ao PC4 (0,65) e que não foi selecionado em outros componentes.

Os resultados de porcentagem da variação por meio de outra ACP, sem a variável sódio, foram obtidos (Tabela 5). A partir desses resultados e segundo o critério de reter os PCs com autovalores maiores que um e variação acumulada maior a 70%, foram escolhidos os primeiros três componentes, com 73,62 % da variação acumulada.

Tabela 5 - Resultados dos autovalores, variação e variação acumulada das primeiras quatro componentes principais depois de eliminar a variável sódio (Na).

Componente Principal	Autovalores	Variação (%)	Variação acumulada (%)
PC1	4,95	33,00	33,00
PC2	4,28	28,55	61,55
PC3	1,81	12,07	73,62
PC4	0,99	6,61	80,23

Comparando as duas ACPs, decidimos trabalhar com os resultados obtidos pela primeira análise, uma vez que gerou maior variância acumulada (78,33%), em comparação com a variância acumulada de 73,62 % da segunda análise.

Segundo Ferreira (2018), pode-se utilizar as variáveis mais relacionadas com os componentes principais de maior explicação (Tabela 6). O primeiro componente principal (PC1) recebeu maior influência do fósforo (P), saturação de bases (SB), manganês (Mn), matéria orgânica (MO), potencial hidrogeniônico (pH), magnésio (Mg), baseado nos coeficientes de correlação com valores de 0,93, 0,81, 0,76, -0,73, 0,64 e 0,60, respectivamente. O comportamento espacial entre o PC1 e os quatro atributos mais correlacionados pode ser observado na Figura 12. Visualmente o mapa do PC1 foi mais parecido com o atributo fósforo, saturação de bases e matéria orgânica. No caso deste último atributo, o comportamento das cores foi o contrário devido ao coeficiente de correlação negativo.

No segundo componente (PC2) foi encontrada uma influência negativa das variáveis ferro (Fe), fósforo (P), saturação de bases (SB), areia, Zinco (Zn) e influência positiva dos atributos capacidade de troca catiônica (CIC), matéria orgânica (MO) e argila (valores de -0,81, -0,74, -0,59, -0,52, -0,51, 0,73, 0,61 e 0,53, respectivamente). Na Figura 13, encontra-se o mapa de distribuição espacial da componente dois com os quatro atributos mais correlacionados. O mapa do PC2 foi mais semelhante aos atributos P, CIC e MO.

O componente 3 apresentou correlações altas com os valores dos atributos fósforo (0,65), zinco (0,59) e saturação de bases (0,57). As variáveis fósforo (0,95), saturação de bases (0,72), potencial hidrogeniônico (0,51) e matéria orgânica (-0,51) tiveram correlações altas com o componente principal 4. Na Figura 14, encontram-se os mapas de distribuição espacial da componente 3 com os atributos mais relacionados. Neste caso, a variável mais parecida com a CP3 foi o fósforo e as outras duas variáveis do solo não foram semelhantes. O mapa da componente 4 com os mapas das quatro variáveis mais correlacionadas está apresentado na Figura 15. O mapa do PC4 é parecido com o mapa do PC1 (Figura 12), pois três dos coeficientes de correlação altos do PC4 foram também altos no PC1 (P, SB e MO).

Correlações médias foram obtidas para 4 variáveis: 0,48 do cálcio com PC1, -0,47 do cobre com PC2, -0,36 do silte com PC3, e -0,30 de sódio com PC4. O boro foi o único atributo do solo com correlações baixas com todos os componentes, sendo o maior coeficiente de correlação (de 0,29) com o PC2.

Tabela 6 - Autovetores e correlações entre cada componente principal e os atributos do solo, resultados obtidos a partir da ACP.

Atributo	Autovetores				Correlações			
	PC1	PC2	PC3	PC4	PC1	PC2	PC3	PC4
pH	0,32	-0,10	0,20	0,15	0,64	-0,39	0,41	0,51
P	0,32	-0,18	0,08	0,32	0,93	-0,74	0,65	0,95
Cu	0,16	-0,24	-0,48	-0,22	0,35	-0,47	-0,39	0,12
Zn	0,13	-0,21	0,45	0,02	0,41	-0,51	0,59	0,49
Fe	0,16	-0,38	-0,25	-0,12	0,49	-0,81	-0,08	0,41
Mn	0,37	-0,06	-0,03	-0,29	0,76	-0,32	0,08	0,32
B	0,02	0,22	-0,36	0,21	-0,02	0,29	-0,19	-0,07
MO	-0,34	0,22	0,06	0,12	-0,73	0,61	-0,22	-0,51
CIC	0,01	0,43	0,07	0,07	-0,10	0,73	-0,02	-0,23
Ca	0,26	0,28	0,19	0,22	0,48	0,26	0,40	0,28
Mg	0,34	0,23	0,05	0,10	0,60	0,14	0,29	0,29
Na	0,03	0,05	0,35	-0,65	-0,07	0,19	0,01	-0,30
SB	0,36	-0,15	0,13	0,21	0,81	-0,59	0,57	0,72
Argila	0,20	0,32	0,03	-0,33	0,20	0,53	-0,09	-0,25
Silte	0,23	0,21	-0,34	0,03	0,34	0,27	-0,36	0,00
Areia	-0,25	-0,34	0,16	0,17	-0,32	-0,52	0,25	0,17

pH: potencial hidrogeniônico, P: fósforo, Cu: cobre, Zn: zinco, Fe: ferro, Mn: manganês, B: boro, MO: matéria orgânica, CIC: capacidade de troca catiônica, Ca: cálcio, Mg: magnésio, Na: sódio, SB: saturação de bases. PC: componente principal.

Figura 12 - Mapas de distribuição espacial do primeiro componente principal e das primeiras quatro variáveis mais correlacionadas: fósforo, saturação de bases, manganês e matéria orgânica.

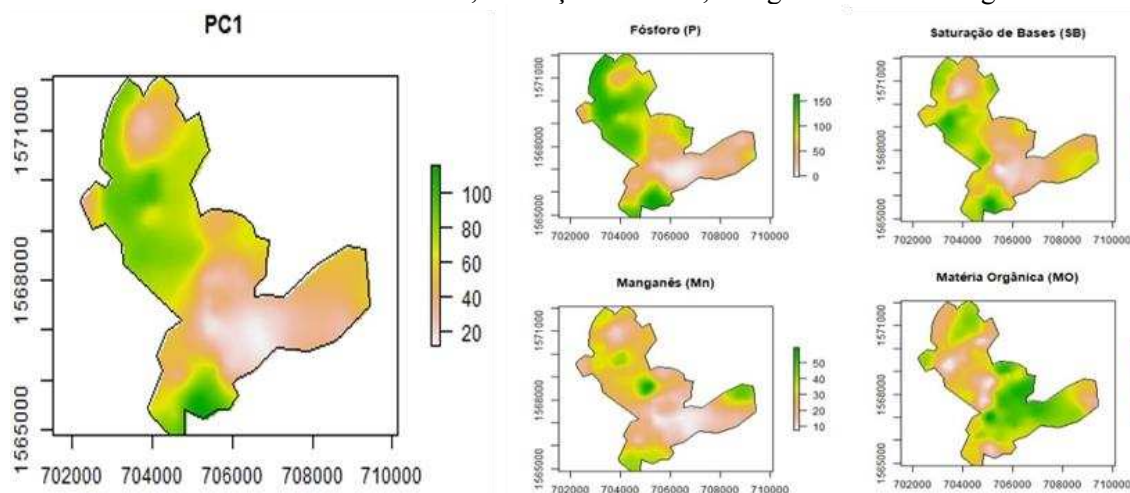


Figura 13 - Mapas de distribuição espacial do componente principal 2 e das primeiras quatro variáveis mais correlacionadas: ferro, fósforo, capacidade de troca catiônica e matéria orgânica.

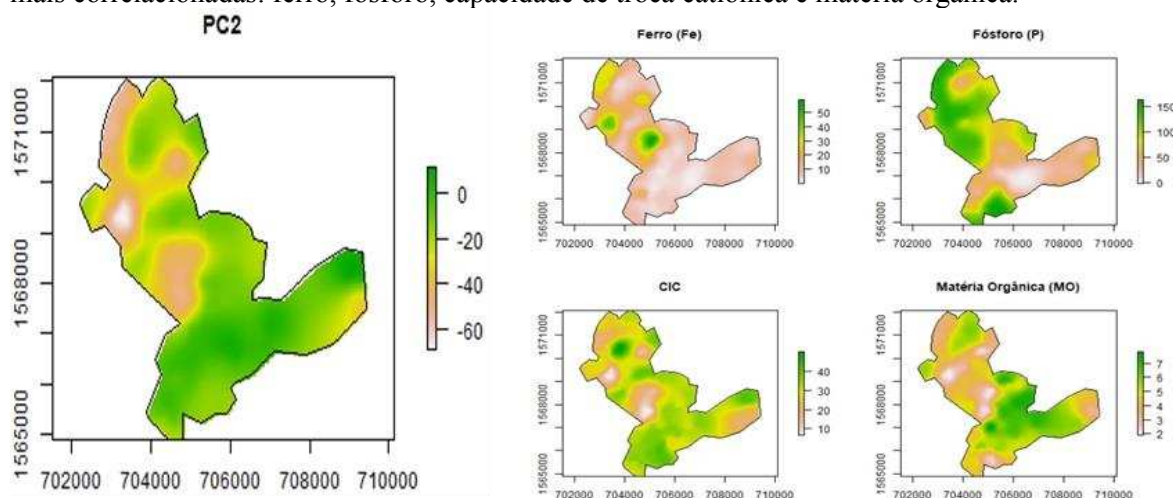


Figura 14 - Mapas de distribuição espacial do componente principal 3 com as variáveis mais correlacionadas, com coeficientes maiores a 0,50: fósforo, zinco e saturação de bases.

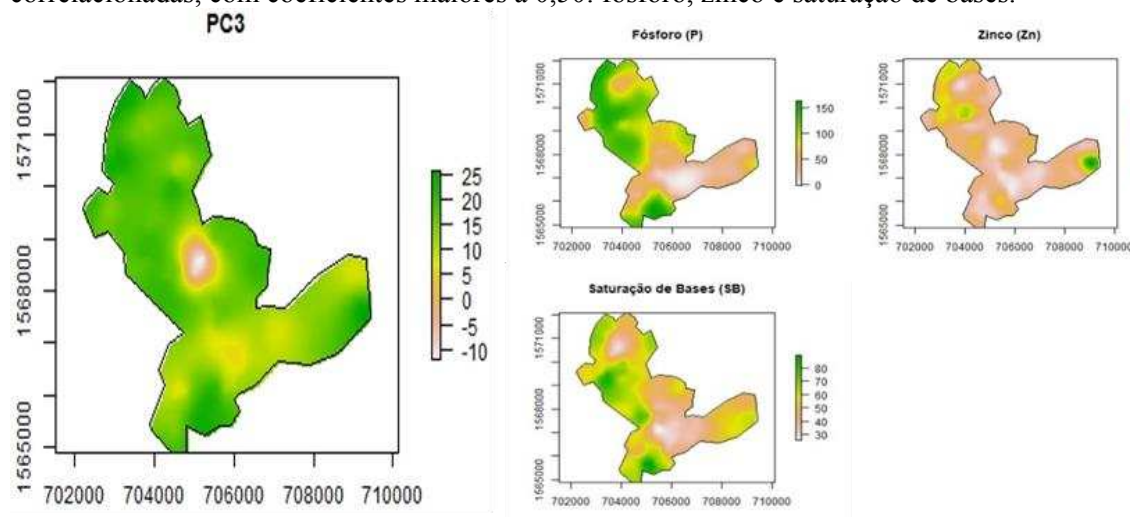
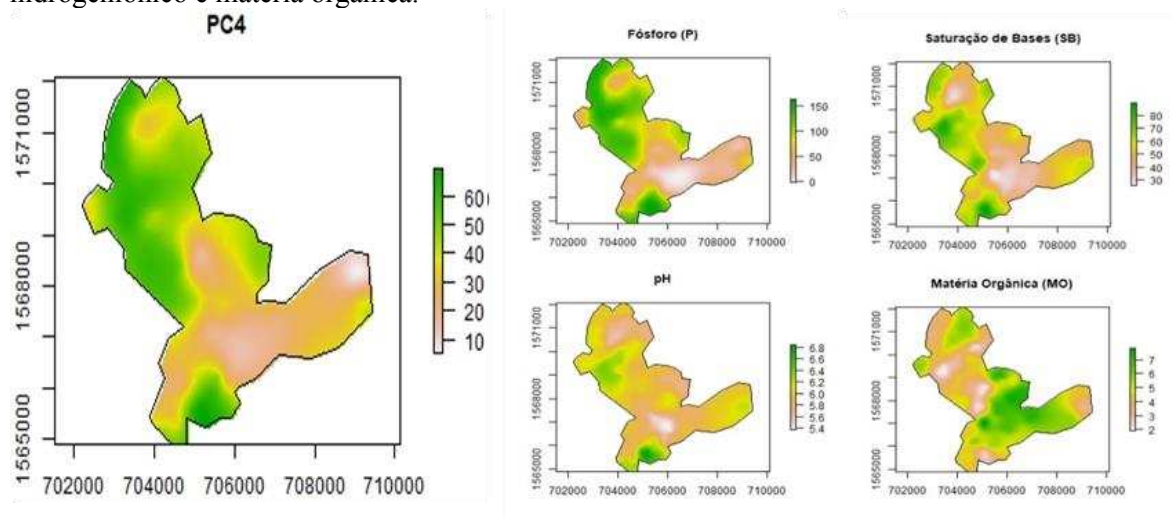


Figura 15 - Mapas de distribuição espacial do componente principal 4 com as variáveis mais correlacionadas, com coeficientes maiores a 0,50 em módulo: fósforo, saturação de bases, potencial hidrogeniônico e matéria orgânica.



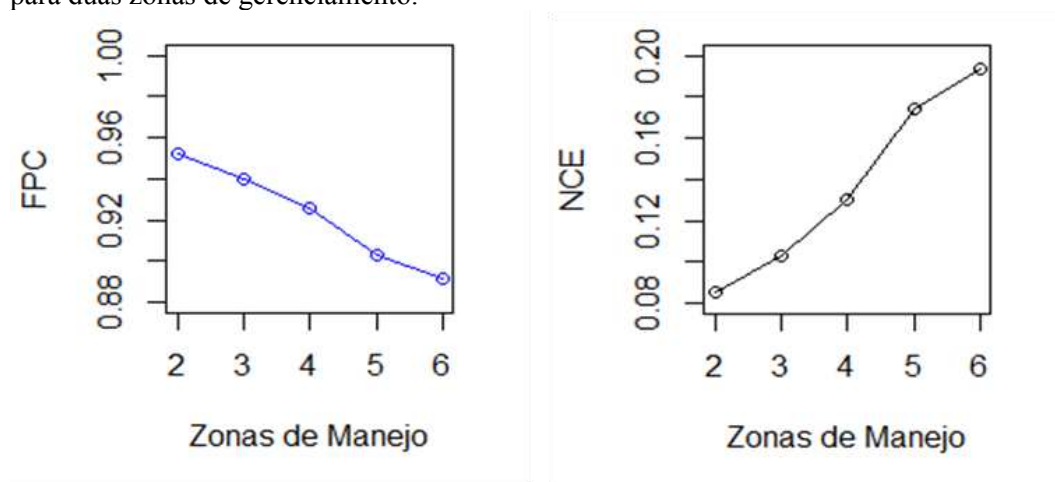
4.5 Análise de agrupamento – Algoritmo Fuzzy K-means

Para definir as ZM com base nos quatro componentes principais, foram importados os valores de escores para o algoritmo Fuzzy K-means. Foram considerados seis *clusters* como o número máximo de zonas de gerenciamento. Para determinar o número ideal de *clusters*, os índices de coeficiente de partição difusa (FPC) e de entropia de classificação normalizada (NCE) foram utilizados. Os resultados dos índices estão apresentados na Tabela 7 e graficamente na Figura 16. Os índices FPC e NCE indicaram que o número ótimo de zonas de manejo foi de duas, com o maior FPC (0,951) e menor NCE (0,090), o que indica maior probabilidade de pertencer a um cluster e não aos demais, além de maior organização de grupos (HE, 2018; TRIPATHI et al., 2015).

Tabela 7 - Coeficiente de partição difusa (FPC) e entropia de partição normalizada (NCE) em relação ao número de *cluster* da área de estudo, com base em quatro componentes principais, calculados a partir dos atributos do solo.

Zonas de manejo	FPC	NCE
2	0,951	0,090
3	0,940	0,103
4	0,926	0,130
5	0,903	0,174
6	0,892	0,193

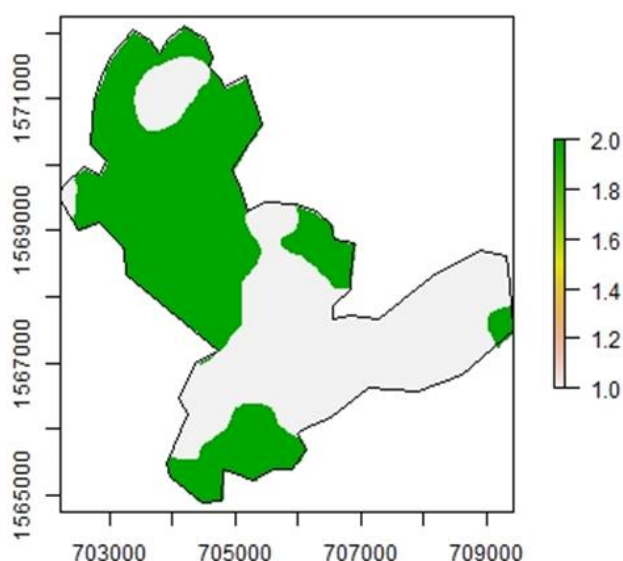
Figura 16 - Índices de validação de zonas de manejo em relação ao número de cluster da área de estudo para duas zonas de gerenciamento.



FPC: coeficiente de partição difusa; NCE: entropia de partição normalizada.

Foi gerado o mapa para as zonas de manejo (ZM) obtidas pelo algoritmo Fuzzy K-means. Cada observação foi classificada em um único grupo (1 ou 2), com a cor branca para a ZM um e com a cor verde para a ZM dois (Figura 17).

Figura 17 - Mapa da delimitação de duas zonas de manejo para os 2 *cluster* obtidos a partir do algoritmo Fuzzy K-means.



Em seguida, foi criada uma nova base de dados para a zona de manejo um (ZM1) e zona de manejo dois (ZM2). A nova base de dados contou com os valores georreferenciados de cada observação das variáveis: potencial hidrogeniônico (pH), fósforo (P), cobre (Cu), zinco (Zn), ferro (Fe), manganês (Mn), boro (B), matéria orgânica (MO), capacidade de troca catiônica (CIC), cálcio (Ca), magnésio (Mg), sódio (Na), saturação de bases (SB), argila, silte e areia. Na Tabela 8, são apresentadas as medidas descritivas para os dados georreferenciados da ZM1 e ZM2.

Valores da média e mediana para cada atributo foram calculadas. Para tornar a comparação das zonas de manejo objetiva, foram obtidos o desvio padrão e o coeficiente de variação, expresso em porcentagem (CV %). Descritivamente, foram comparados os coeficientes de variação para as duas ZM. Na primeira zona de manejo, os resultados do CVs obtidos foram maiores para os atributos pH, P, Cu, Zn, Mn, B, Mg, Na, SB, argila, silte e areia, em comparação à zona de manejo 2.

Tabela 8 - Medidas descritivas dos dados gerados para duas zonas de manejo delimitadas pelo algoritmo Fuzzy K-means.

Atributo	ZM	Média	Mediana	DP	CV (%)	Assimetria	Curtose
pH	1	5,91	5,96	0,14	2,41	-1,27	1,53
	2	6,01	5,98	0,13	2,15	1,76	7,60
P (ppm)	1	42,86	44,61	18,63	43,48	-0,39	-0,35
	2	94,98	82,74	22,76	23,96	1,10	0,08
Cu (ppm)	1	0,41	0,40	0,37	88,99	4,32	30,55
	2	0,70	0,61	0,50	72,06	5,56	42,49
Zn (ppm)	1	5,92	6,22	1,61	27,20	1,25	7,35
	2	7,16	6,85	1,60	22,37	2,84	15,40
Fe (ppm)	1	7,41	7,32	3,85	52,00	2,12	11,20
	2	13,03	11,49	6,79	52,11	2,72	9,50
Mn (ppm)	1	20,17	17,78	9,05	44,87	1,10	0,53
	2	26,69	25,26	4,94	18,50	1,50	4,86
B (ppm)	1	0,83	0,84	0,09	10,36	0,01	0,20
	2	0,84	0,84	0,07	7,81	-0,34	3,96
MO (%)	1	5,59	5,53	0,87	15,62	0,17	-0,70
	2	4,49	4,62	0,71	15,87	-0,38	2,31
CIC (meq/100g)	1	30,74	30,64	4,37	14,22	-0,19	1,63
	2	29,02	29,43	4,71	16,24	-1,28	5,15
Ca (meq/100g)	1	10,64	10,66	1,78	16,75	0,07	0,26
	2	12,01	11,63	2,46	20,52	1,47	7,03
Mg (meq/100g)	1	2,56	2,37	0,70	27,44	0,82	0,91
	2	3,47	3,30	0,92	26,39	0,70	1,29
Na (meq/100g)	1	0,35	0,32	0,12	34,36	1,01	0,55
	2	0,33	0,33	0,07	21,05	0,42	2,43
SB (%)	1	46,23	47,52	7,93	17,16	-0,39	-0,81
	2	57,22	54,41	6,58	11,50	1,36	3,54
Argila (%)	1	14,20	12,63	5,63	39,68	0,80	-0,21
	2	14,99	15,45	4,20	28,03	-0,17	-0,76
Silte (%)	1	31,66	32,19	3,59	11,34	-0,42	0,65
	2	32,55	32,25	3,24	9,95	0,97	3,71
Areia (%)	1	54,43	55,42	7,92	14,55	-0,21	-0,30
	2	53,45	53,84	5,64	10,56	-0,11	1,40

pH: potencial hidrogeniônico, P: fósforo, Cu: cobre, Zn: zinco, Fe: ferro, Mn: manganês, B: boro, MO: matéria orgânica, CIC: capacidade de troca catiônica, Ca: cálcio, Mg: magnésio, Na: sódio, SB: saturação de bases, ZM: zona de manejo, Média: média aritmética, DP: desvio padrão, CV (%): coeficiente de variação em porcentagem.

Para conhecer a diferença entre duas médias independentes de uma distribuição normal, podem ser realizados o teste de hipótese de t de Student ou a análise de variância (ANOVA) com comparação de médias pós-ANOVA. Entretanto, os dados para ZM1 e ZM2 não apresentaram distribuição normal, ainda que fazendo transformações logarítmicas e raiz

quadrada. A maioria dos valores de forma (assimetria e curtose) dos atributos do solo foram diferentes descritivamente e, em alguns casos, até com sinais diferentes. Portanto, tais análises paramétricas não foram realizadas. De acordo com Turcios (2015), não foi possível realizar o teste de hipótese não paramétrico Mann-Whitney, que realiza comparação entre medianas de duas amostras independentes, pois os atributos não apresentaram valores similares de viés, curtose e dispersão entre uma zona de manejo e outra.

Portanto, para comparar os coeficientes de variação de cada atributo entre os valores originais e as zonas de manejo obtidas pelo algoritmo Fuzzy K-means, foram resumidas as medidas descritivas da média aritmética, desvio padrão e coeficiente de variação da área de estudo total e das duas zonas de manejo (Tabela 9). A comparação desses valores para todos os atributos do solo indicou que a variabilidade das zonas de manejo foi menor em relação à área geral do estudo, cumprindo o objetivo do algoritmo de minimizar a variabilidade dentro dos grupos, produzindo grupos homogêneos.

Tabela 9 - Comparação entre os valores da média aritmética, desvio padrão e coeficiente de variação para a área total e para as duas zonas de manejo delimitadas pelo algoritmo Fuzzy K-means dos atributos do solo em estudo.

Atributo	Média aritmética			Desvio padrão			CV%		
	Total	ZM1	ZM2	Total	ZM1	ZM2	Total	ZM1	ZM2
pH	5,97	5,91	6,01	0,25	0,14	0,13	4,19	2,41	2,15
P (ppm)	80,42	42,86	94,98	57,55	18,63	22,76	71,56	43,48	23,96
Cu (ppm)	0,61	0,41	0,70	0,91	0,37	0,50	147,86	88,99	72,06
Zn (ppm)	6,70	5,92	7,16	3,73	1,61	1,60	55,63	27,20	22,37
Fe (ppm)	11,59	7,41	13,03	11,94	3,85	6,79	103,06	52,00	52,11
Mn (ppm)	24,08	20,17	26,69	11,65	9,05	4,94	48,38	44,87	18,50
B (ppm)	0,85	0,83	0,84	0,21	0,09	0,07	24,37	10,36	7,81
MO (%)	4,85	5,59	4,49	1,57	0,87	0,71	32,33	15,62	15,87
CIC (meq/100g)	29,61	30,74	29,02	8,31	4,37	4,71	28,08	14,22	16,24
Ca (meq/100g)	11,37	10,64	12,01	4,29	1,78	2,46	37,70	16,75	20,52
Mg (meq/100g)	2,99	2,56	3,47	1,63	0,70	0,92	54,71	27,44	26,39
Na (meq/100g)	0,30	0,35	0,33	0,12	0,12	0,07	41,06	34,36	21,05
SB (%)	52,90	46,23	57,22	14,65	7,93	6,58	27,69	17,16	11,50
Argila (%)	13,41	14,20	14,99	6,45	5,63	4,20	48,09	39,68	28,03
Silte (%)	32,77	31,66	32,55	6,49	3,59	3,24	19,81	11,34	9,95
Areia (%)	53,82	54,43	53,45	11,32	7,92	5,64	21,03	14,55	10,56

pH: potencial hidrogeniônico, P: fósforo, Cu: cobre, Zn: zinco, Fe: ferro, Mn: manganês, B: boro, MO: matéria orgânica, CIC: capacidade de troca catiônica, Ca: cálcio, Mg: magnésio, Na: sódio, SB: saturação de bases, CV: coeficiente de variação em %, Total: área total dos dados, ZM1: zona de manejo 1, ZM2: zona de manejo 2.

As faixas de valores adequados dos atributos de solos franco arenosos (característicos das áreas de produção de cana-de-açúcar da Guatemala) estão apresentadas na Tabela 10 (Pérez, 2013). Na Tabela 10, também estão apresentadas as médias aritméticas dos atributos: potencial hidrogeniônico (pH), matéria orgânica (MO), fósforo (P), cálcio (Ca), magnésio (Mg), boro (B), ferro (Fe), manganês (Mn), cobre (Cu), zinco (Zn) e porcentagem de sódio intercambiável (Na PSI %) para as duas zonas de manejo obtidas no trabalho.

Tabela 10 - Amplitude de valores adequados dos atributos dos solos característicos das áreas de produção de cana-de-açúcar da região sul da Guatemala segundo Pérez (2013), com conteúdo menor a 35 % de argila e valores dos atributos georreferenciados para as duas zonas de manejo.

Atributo	Adequado	ZM1	ZM2
pH	5,5-7,2	5,91	6,01
MO (%)	3,0-5,0	5,59	4,49
P (ppm)	10,0-30,0	42,86	94,98
Ca (meq/100g)	4,0-30,0	10,64	12,01
Mg (meq/100g)	1,0-12,0	2,56	3,47
B (ppm)	0,5-15,0	0,83	0,84
Fe (ppm)	5,0-25,0	7,41	13,03
Mn (ppm)	2,0-20,0	20,17	26,69
Cu (ppm)	0,5-10,0	0,41	0,7
Zn (ppm)	1,0-15,0	5,92	7,16
Na (PSI %)	<15,0	1,14	1,14

pH: potencial hidrogeniônico, MO: matéria orgânica, P: fósforo, Ca: cálcio, Mg: magnésio, B: boro, Fe: ferro, Mn: manganês, Cu: cobre, Zn: zinco, Na (PSI %): porcentagem de sódio intercambiável, ZM1: zona de manejo um, ZM2: zona de manejo dois. Fonte: Os dados da coluna 'Adequado' foram extraídos de Pérez (2013). Os demais dados foram elaborados pela autora.

As variáveis pH, Ca, Mg, B, Fe, Mn, Zn e Na apresentaram valores adequados, enquanto a matéria orgânica pode ser classificada como alta na ZM1 e adequada na ZM2. Por sua vez, os atributos fósforo e manganês apresentaram valores altos e característicos para as duas zonas de manejo. O cobre na ZM1 foi o único nutriente com valores baixos.

De acordo com Partelli et al. (2018), a saturação de bases (SB) é adequada quando é maior que 50%. A saturação de bases na ZM1 foi de 46,23% e na ZM2 foi de 57,22%, o que indica que a ZM1 não apresentou níveis adequados.

Para o crescimento e desenvolvimento das plantas, 13 elementos mineiras são essenciais, e podem advir do solo ou pela aplicação de fertilizantes. Destes 13 elementos, 6 são considerados macronutrientes, fazendo referência aos nutrientes que as plantas necessitam em maior quantidade: nitrogênio (N), fósforo (P), potássio (K), cálcio (Ca), magnésio (Mg) e enxofre (S). N, P, e K são classificados com primários, enquanto Ca, Mg e S são macronutrientes secundários. Os micronutrientes que são requeridos em menor quantidade são:

ferro (Fe), manganês (Mn), zinco (Zn), cobre (Cu), boro (B), molibdênio (Mo) e cloro (Cl). Assim, o manejo para a incorporação de fertilizantes depende dos níveis baixos, adequados e altos dos atributos do solo obtidos pelas análises químicas das amostras. Os planos de fertilização também dependem da época de aplicação, localização, textura do solo, irrigação, fonte confiável de dados e custos econômicos.

O solo para as duas zonas de manejo foi classificado como franco-arenoso e um pH em faixas de valores adequados, que influenciam na disponibilidade de nutrientes do solo. O conteúdo da matéria orgânica para a ZM1 foi alto, com 5,59%.

De acordo com Cengicaña (2012), o nitrogênio (N) é encontrado no solo, em sua maior parte, em formas orgânicas (geralmente mais de 95%). Além disso, a resposta a aplicações de N na cana-de-açúcar da Guatemala apresenta alta correlação com os conteúdos da matéria orgânica (MO) do solo. Na Figura 21 (Anexo), é apresentada a relação direta entre a MO e a porcentagem de casos de aumentos da produção, expressa em toneladas de cana por hectares (TCH), devido às aplicações de N na cultura (Cengicaña, 2012). Cengicaña (2012) observou que, em solos com 3 a 5% de MO, houve aumento de 11 a 20% das TCH em 47% dos casos de aplicação de N. Em solos com mais de 5% de MO, por sua vez, o aumento de TCH do solo foi menor que 11% em 100% dos casos. Com base a nesses dados pode-se esperar que na ZM1 sejam necessárias menos aplicações de fertilizantes com N.

Para o macronutriente primário fósforo (P), valores elevados foram obtidos para as duas zonas de manejo, com maior magnitude na zona de manejo 2 (94,98 ppm, enquanto na ZM1 foram encontrados 42,86 ppm). Para níveis altos, não são consideradas as aplicações de P mas, ainda assim, é necessária a mobilidade desse nutriente no solo (CENGICAÑA, 2012).

Os níveis de potássio (K: macronutriente primário), que não foi considerado como parte da análise ACP e nem do algoritmo Fuzzy K-means, foram altos, com 0,52 meq/100g do solo. Segundo Cengicaña (2012), níveis adequados de potássio estão entre 0,25 e 0,38 meq/100g. Entretanto, no nosso estudo o conteúdo do potássio foi considerado alto para a área total de estudo. De acordo com Pérez (2013), o excesso de K aumenta o teor de cinzas no suco da cana-de-açúcar e reduz a recuperação de sacarose, levando a problemas na cristalização e comportamento antagonista na absorção de cálcio e magnésio. No caso de micronutrientes, o manganês foi alto tanto na primeira como na segunda zona de manejo, com 20,17 ppm e 26,69 ppm respectivamente. Para o atributo cobre, foram obtidos valores baixos para a ZM1 (0,41 ppm).

A saturação de bases para a ZM1 foi de 46,23%, sendo recomendável elevar o percentual a valores superiores a 50% ou a valores que sejam os ideais para o cultivo de cana-de-açúcar na fazenda Tehuantepec.

Os resultados mais importantes foram discutidos anteriormente, baseados nos macronutrientes primários. É relevante destacar que a disponibilidade dos nutrientes está relacionada a vários fatores, como solo, planta e meio ambiente. Desta forma, não se deve conduzir o manejo de fertilização baseando-se somente em um critério, como os níveis adequados dos atributos do solo. Tal afirmação é confirmada por Pérez (2013): “A análise do solo é uma ferramenta e seu sucesso dependerá em alto grau de critérios científicos e técnico profissional, buscando obter produtividade e ótima rentabilidade, com uso eficiente de insumos e proteção do meio ambiente”. Outros critérios, como variedade da cana, época de aplicações, tipo de argila, relações catiônicas, análise químico foliar, interação entre nutrientes e condições climáticas, devem ser considerados.

5 CONCLUSÃO

No presente trabalho, foram delineadas zonas de manejo em uma lavoura de cana-de-açúcar (*Saccharum officinarum*), na região sul da Guatemala, utilizando a metodologia espacial e multivariada. Foram aplicadas as análises Geoestatística, análise de componentes principais e o algoritmo de agrupamento Fuzzy K-means.

A Geoestatística foi aplicada para estudar a variabilidade espacial. A análise variográfica determinou que 16 das 17 variáveis do solo mostraram dependências espaciais, o que permitiu que os dados fossem interpolados por meio da técnica de Krigagem para a construção e visualização de mapas de variabilidade espacial com valores preditos que serviram de entrada para análise multivariada.

Os índices de coeficiente de partição difusa (FPC) e de entropia de classificação normalizada (NCE) corroboraram para que o número ótimo fosse de duas zonas de manejo. Embora não tenha sido possível testar estatisticamente a diferença de médias ou medianas entre as zonas de manejo de cada um dos atributos do solo, foram comparados os coeficientes de variação de forma descritiva, detectando valores diferentes de variabilidade em cada zona de manejo, o que pode permitir uma gestão mais específica nos planos de fertilização do solo, a fim de melhorar sua eficiência.

Os resultados demonstraram que a ZM1 e a ZM2 podem ser gerenciadas de forma específica, baseando-se nos atributos fósforo, manganês, cobre e matéria orgânica. Os níveis de matéria orgânica tiveram alta correlação com o teor de nitrogênio, podendo ser utilizados como base para aplicações de fertilizantes nitrogenados.

Portanto, o presente trabalho comprova que a delimitação de zonas de manejo é uma alternativa promissora para otimizar o uso de fertilizantes no solo. Tal delimitação pode ser eficientemente realizada pelas técnicas multivariadas de análise de componentes principais e pelo algoritmo Fuzzy K-means, utilizando dados obtidos por Krigagem.

REFERÊNCIAS BIBLIOGRÁFICAS

BALBOA, W. J. **Modelamiento y Co-Simulación de Leyes de Cobre Total y Soluble**. Santiago de Chile, 2012.

BEHERA, S. K.; MATHUR, R. K.; SHUKLA, A. K.; SURESH, K.; PRAKASH, C. Spatial variability of soil properties and delineation of soil management zones of oil palm plantations grown in a hot and humid tropical region of southern India. **Catena**, [S. l.], v. 165, n. Jun. 2017, p. 251–259, 2018. DOI: 10.1016/j.catena.2018.02.008. Disponível em: <https://doi.org/10.1016/j.catena.2018.02.008>. Acesso em: 11 Oct. 2020.

BERNARDI, A. C. de C.; NAIME, J. de M.; DE RESENDE, A. V.; BASSOI, L. H.; INAMASU, R. Y. **Agricultura de precisão**: Resultados de um novo olhar. 1a. ed. Brasília: Embrapa Instrumentação-Livro técnico (INFOTECA-E), 2014.

BONGIOVANNI, R.; MANTOVANI, E. C.; BEST, S.; ROEL, A. **Agricultura de precisión**: Integrando conocimientos para una agricultura moderna y sustentable. Montevideo: PROCISUR/IICA, 2006.

CAMBARDELLA, C. A.; MOORMAN, T. B.; NOVAK, J. M.; PARKIN, T. B.; KARLEN, D. L.; TURCO, R. F.; KONOPKA, A. E. Field-Scale Variability of Soil Properties in Central Iowa Soils. **Soil Science Society of America Journal**, [S. l.], v. 58, n. 5, p. 1501–1511, 1994. DOI: 10.2136/sssaj1994.03615995005800050033x.

CEBECI, Z.; YILDIZ, F.; TUNA, A.; CEBECI, C.; ONDER, H. **Probabilistic and Possibilistic Cluster Analysis Package ppclust**. [s.l: s.n.], 2018. DOI: 10.1109/91.227387.

CENGICAÑA. **El Cultivo de la Caña de Azúcar en Guatemala**. Guatemala, 2012.

COHEN, J. **Statistical Power Analysis for the Behavioral Sciences**. 3a. ed. New York: Lawrence Erlbaum Associates, 1988.

CÓRDOBA, M. A. **Herramientas estadísticas para el monitoreo y uso de la variabilidad espacial del rendimiento y propiedades de suelo intralote**. 2014. Tese (Doutorado em Ciências Agropecuárias) - Facultad de Ciencias Agropecuarias, Universidad Nacional de Córdoba, [S. l.], 2014.

FERREIRA, D. **Estatística Multivariada**. 3a. ed. Lavras: UFLA, 2018.

FRIDGEN, J. J.; KITCHEN, N. R.; SUDDUTH, K. A.; DRUMMOND, S. T.; WIEBOLD, W. J.; FRAISSE, C. W. Management Zone Analyst (MZA) Software for Subfield Management Zone Delineation. **Agronomy Journal**, [S. l.], p. 9, 2004. DOI: <https://doi.org/10.2134/agronj2004.1000>. Disponível em: <https://acess.onlinelibrary.wiley.com/doi/epdf/10.2134/agronj2004.1000>. Acesso em: 20 Set. 2019.

FU, Q.; WANG, Z.; JIANG, Q. Delineating soil nutrient management zones based on fuzzy clustering optimized by PSO. **Mathematical and Computer Modelling**, [S. l.], v. 51, n. 11–12, p. 1299–1305, 2010. DOI: 10.1016/j.mcm.2009.10.034. Disponível em: <http://dx.doi.org/10.1016/j.mcm.2009.10.034>. Acesso em: 05 Ago. 2020.

GAVIOLI, A.; DE SOUZA, E. G.; BAZZI, C. L.; GUEDES, L. P. C.; SCHENATTO, K. Optimization of management zone delineation by using spatial principal components. **Computers and Electronics in Agriculture**, [S. l.], v. 127, p. 302–310, 2016. DOI: 10.1016/j.compag.2016.06.029. Disponível em: <http://dx.doi.org/10.1016/j.compag.2016.06.029>. Acesso em: 13 Set. 2020.

GAVIOLI, A.; DE SOUZA, E. G.; BAZZI, C. L.; SCHENATTO, K.; BETZEK, N. M. Identification of management zones in precision agriculture: An evaluation of alternative cluster analysis methods. **Biosystems Engineering**, [S. l.], v. 181, p. 86–102, 2019. DOI: 10.1016/j.biosystemseng.2019.02.019. Disponível em: <https://doi.org/10.1016/j.biosystemseng.2019.02.019>. Acesso em: 15 Set. 2020.

GILI, A.; ÁLVAREZ, C.; BAGNATO, R.; NOELLEMEYER, E. Comparison of three methods for delineating management zones for site-specific crop management. **Computers and Electronics in Agriculture**, [S. l.], v. 139, p. 213–223, 2017. DOI: 10.1016/j.compag.2017.05.022. Disponível em: <http://dx.doi.org/10.1016/j.compag.2017.05.022>. Acesso em: 15 Set. 2020.

HE, Y. **Data Analysis for the Mobile Application of the selfBACK Decision Support System**. Norwegian University of Science and Technology, [S. l.], p. 86, 2018.

HEANO, R. G. **Introducción a la Geoestadística**. Teoría y aplicación. Bogotá: Departamento de estadística. UNA., 2002.

HENRÍQUEZ, C.; MÉNDEZ, J. C.; MASÍS, R. Interpolación de variables de fertilidad de suelo mediante el análisis Kriging y su validación. **Agronomía Costarricense**, [S. l.], v. 37, n. 2, p. 71–82, 2013. DOI: 10.15517/rac.v37i2.12763.

INIA. **Agricultura de precisión. Introducción al manejo sitio-específico**. Chillán, Chile, 1999.

JAIN, A. K. Data clustering: 50 years beyond K-means. **Pattern Recognition Letters**, [S. l.], v. 31, n. 8, p. 651–666, 2010. DOI: 10.1016/j.patrec.2009.09.011. Disponível em: <http://dx.doi.org/10.1016/j.patrec.2009.09.011>. Acesso em: 21 Out. 2020.

JOLLIFFE, I. T. **Principal component analysis**. New York: Springer-Verlag, 2002. DOI: 10.1007/b98835.

LI, Y.; SHI, Z.; LI, F.; LI, H. Y. Delineation of site-specific management zones using fuzzy clustering analysis in a coastal saline land. **Computers and Electronics in Agriculture**, [S. l.], v. 56, n. 2, p. 174–186, 2007. DOI: 10.1016/j.compag.2007.01.013. Acesso em: 28 Jun. 2020.

LIKAS, A.; VLASSIS, N.; VERBEEK, J. J. **O algoritmo global de agrupamento k- significa** **Palavras-chave**. [S. l.], p. 1–19, 2020.

LOUZADA, J. M. **Teste de normalidade multivariada em geoestatística utilizando Bootstrap**. Lavras: UFLA, 2011.

MATHERON, G. **Le krigeage univesel. Cahiers du Centre de Morphologie Mathématique**. vol. 1 ed. Paris: Ecole des Mines de Paris, Fontainebleau., 1970.

MEYER, D.; DIMITRIADOU, E.; HORNIK, K.; WEINGESSEL, A.; FRIEDRICH, L.; CHANG, C.; LIN, C. **Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. Version 1.7-6.** [s.l: s.n.], 2021. Disponível em: <https://cran.r-project.org/web/packages/e1071/e1071.pdf>. Acesso: 06 Jul. 2020.

MOLIN, J. P.; DE CASTRO, C. N. Establishing management zones using soil electrical conductivity and other soil properties by the fuzzy clustering technique. **Scientia Agricola**, [S. l.], v. 65, n. 6, p. 567–573, 2008. DOI: 10.1590/S0103-90162008000600001.

MORAL, F. J.; TERRÓN, J. M.; REBOLLO, F. J. Site-specific management zones based on the Rasch model and geostatistical techniques. **Computers and Electronics in Agriculture**, [S. l.], v. 75, n. 2, p. 223–230, 2011. DOI: 10.1016/j.compag.2010.10.014. Disponível em: <http://dx.doi.org/10.1016/j.compag.2010.10.014>. Acesso em: 20 Mar. 2021.

MORAL, F. J.; TERRÓN, J. M.; SILVA, J. R. Marque. Da. Delineation of management zones using mobile measurements of soil apparent electrical conductivity and multivariate geostatistical techniques. **Soil and Tillage Research**, [S. l.], v. 106, n. 2, p. 335–343, 2010. DOI: 10.1016/j.still.2009.12.002.

OHANA-LEVI, N.; BAHAT, I.; PEETERS, A.; SHTEIN, A.; NETZER, Y.; COHEN, Y.; BEN-GAL, A. A weighted multivariate spatial clustering model to determine irrigation management zones. **Computers and Electronics in Agriculture**, [S. l.], v. 162, n. February, p. 719–731, 2019. DOI: 10.1016/j.compag.2019.05.012.

OLDONI, H.; TERRA, V. S. S.; TIMM, L. C.; JÚNIOR, C. R.; MONTEIRO, A. B. Delineation of management zones in a peach orchard using multivariate and geostatistical analyses. **Soil and Tillage Research**, [S. l.], v. 191, n. January, p. 1–10, 2019. DOI: 10.1016/j.still.2019.03.008. Disponível em: <https://doi.org/10.1016/j.still.2019.03.008>.

ORTEGA, R. A.; SANTIBÁÑEZ, O. A. Determination of management zones in corn (*Zea mays* L.) based on soil fertility. **Computers and Electronics in Agriculture**, [S. l.], v. 58, n. 1, p. 49–59, 2007. DOI: 10.1016/j.compag.2006.12.011.

PARTELLI, F. L.; EVANGELISTA, C. R.; CAVALANCATI, A.; GONTIJO, I. Orgánica y convencional - Properties of the fertility of a cane soil under different types of organic and conventional management. **Cultivos Tropicales (instituto nacional de ciencias agrícolas)**, [S. l.], v. 39, n. 4, p. 13–20, 2018.

PERALTA, N.; COSTA, J. L. Delineation of management zones with soil apparent electrical conductivity to improve nutrient management. **Computers and Electronics in Agriculture**, [S. l.], v. 99, p. 218–226, 2013. DOI: 10.1016/j.compag.2013.09.014.

PÉREZ, O. **Fertilización y nutrición del cultivo de caña de azúcar**. Cengincña, 2013.

RIBEIRO JUNIOR, P. J.; DIGGLE, P. J.; CHRISTENSEN, O.; SCHLATHER, M.; BIVAND, R.; RIPLEY, B. **geoR: Analysis of Geostatistical Data**. [S. l.], p. 155, 2020. DOI: 10.1007/978-0-387-48536-2>.License. Disponível em: <https://cran.r-project.org/package=geoR>. Acesso em: 20 Mar. 2021.

ROSA, L. M. F. **Estudos sobre a influência de afirmações populares na geoestatística clássica**. 2017. Tese (Doutorado em Estatística Aplicada e Biometria) - Departamento de Estatística, Universidade Federal de Viçosa, Viçosa, MG, 2017.

SANTOS, G. R. **Hierarquização geométrica dos preditores geoestatísticos**. 2010. Tese (Doutorado em Estatística e Experimentação Agropecuária) - Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, Universidade Federal de Lavras, Lavras, MG, 2010.

SOUZA, Z. M.; CERRI, D. G. P.; COLET, M. J.; RODRIGUES, L. H. A.; MAGALHÃES, P. S. G.; MANDONI, R. J. A. Análise dos atributos do solo e da produtividade da cultura de cana-de-açúcar com o uso da geoestatística e árvore de decisão. **Ciência Rural**, [S. l.], v. 40, n. 4, p. 840–847, 2010. DOI: 10.1590/s0103-84782010005000048.

R CORE TEAM. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing Viena, Austria, 2020. Disponível em: <https://www.r-project.org/>. Acesso em: 20 Fev. 2021.

TRIPATHI, R. et al. Delineation of soil management zones for a rice cultivated area in eastern India using fuzzy clustering. **Catena**, [S. l.], v. 133, p. 128–136, 2015. DOI: 10.1016/j.catena.2015.05.009. Disponível em: <http://dx.doi.org/10.1016/j.catena.2015.05.009>. Acesso em: 20 Fev. 2021.

TURCIOS, R. A. S. Prueba de Wilcoxon-Mann-Whitney: mitos y realidades. **Rev Mex Endocrinol Metab Nutr**, [S. l.], v. 2, p. 18–21, 2015. Disponível em: www.endocrinologia.org.mx. Acesso em: 07 Abr. 2020.

VIEIRA, S. R. **Geoestatística em estudos de variabilidade espacial do solo**. Tópicos em ed. Viçosa - MG: V1, 2000.

WEBSTER, R.; OLIVER, M. A. **Geostatistics for Environmental Scientists**. 2. ed. Chichester, 2007.

XIN-ZHONG, W.; GUO-SHUN, L.; HONG-CHAO, H.; ZHEN-HAI, W.; QING-HUA, L.; XU-FENG, L.; WEI-HONG, H.; YAN-TAO, L.. Determination of management zones for a tobacco field based on soil fertility. **Computers and Electronics in Agriculture**, [S. l.], v. 65, n. 2, p. 168–175, 2009. DOI: 10.1016/j.compag.2008.08.008.

APÊNDICE

Figura 18 - Análise exploratória gráfica dos atributos do solo. a) Potencial hidrogeniônico, b) fósforo, c) cobre, d) zinco, e) ferro, f) manganês, g) boro, h) matéria orgânica.

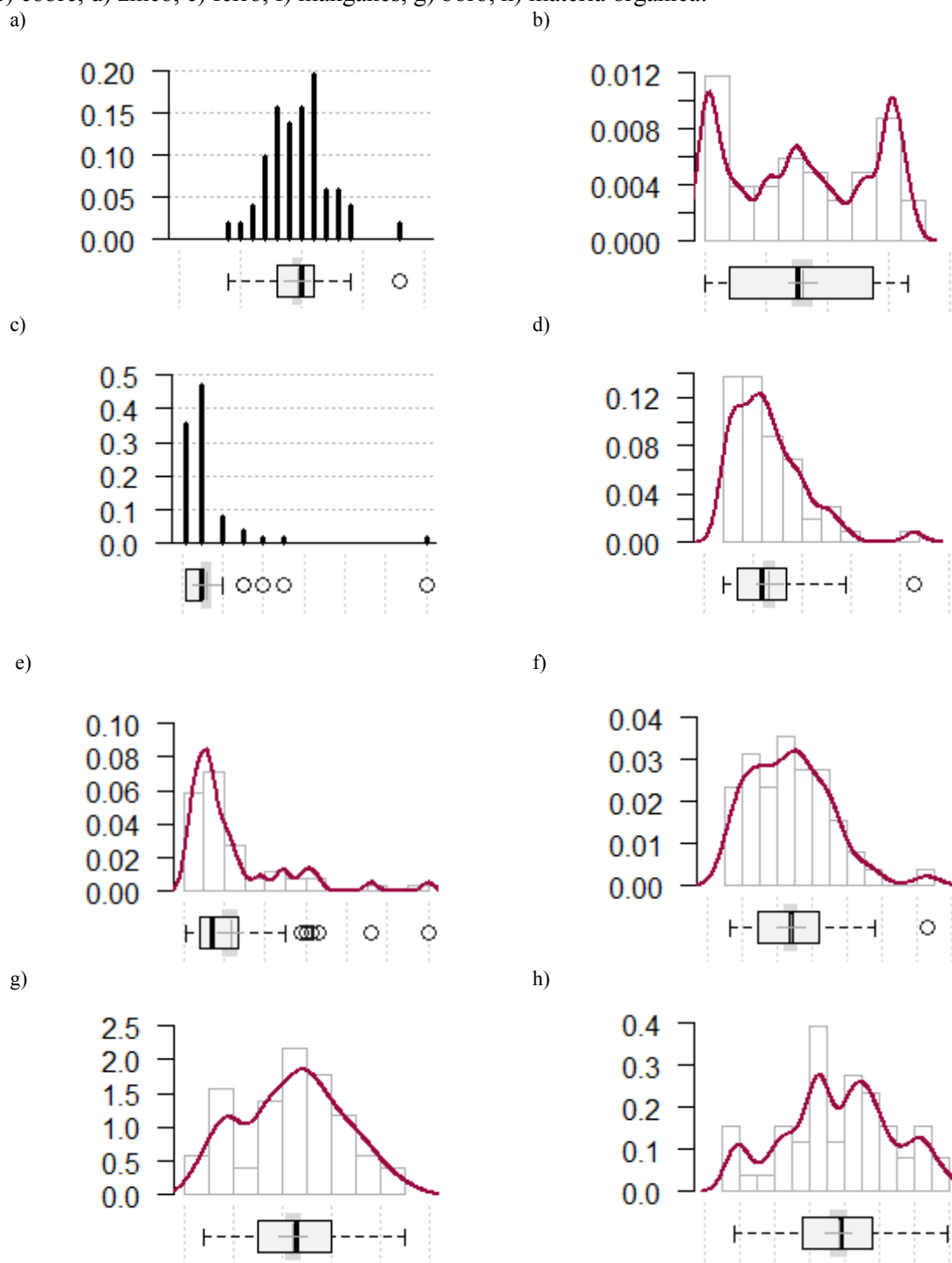
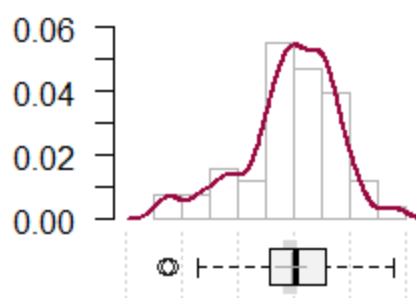
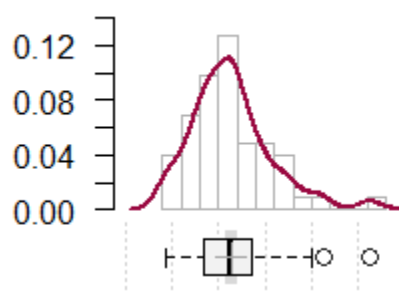


Figura 19 - Análise exploratória gráfica dos atributos do solo. a) Capacidade de troca catiônica, b) cálcio, c) magnésio, d) sódio, e) potássio, f) saturação de bases, g) argila, h) silte.

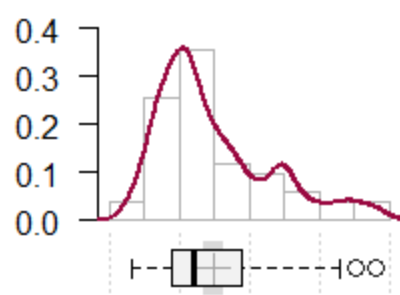
a)



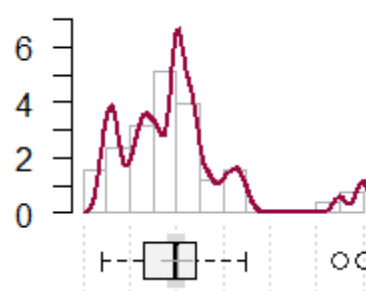
b)



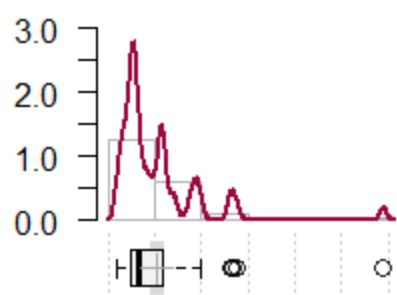
c)



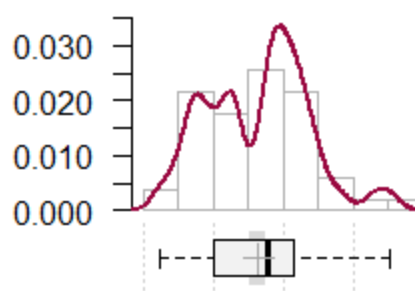
d)



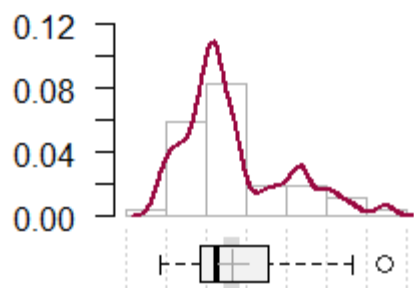
e)



f)



g)



h)

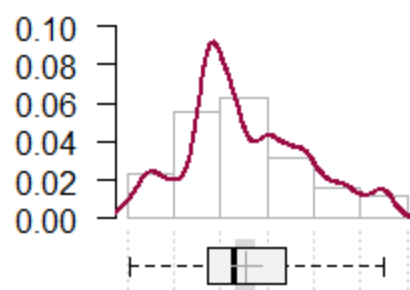


Figura 20 - Análise exploratória gráfica do atributo areia.

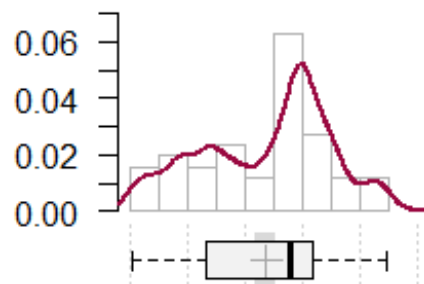


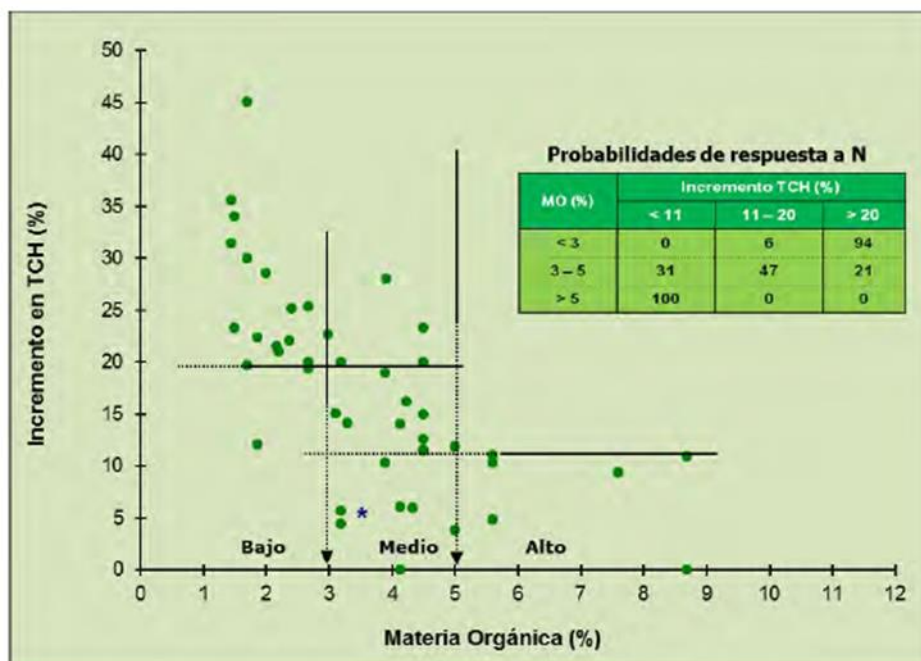
Tabela 11 - Matriz de correlação linear de Pearson.

CIC																0,70*
Mg															0,51*	0,74*
Ca													0,74*	0,63*	0,70*	
B												0,20*	0,35*	0,32*	0,26*	
Na										-0,10	-0,03	0,10	0,05	0,19*		
Silte									-0,15	0,26*	0,46*	0,55*	0,39*	0,53*		
Mn								0,52*	0,07	-0,02	0,32*	0,43*	-0,09	0,25*		
K							0,33*	0,31*	0,00	0,21*	0,29*	0,32*	0,24*	0,24*		
SB						0,18*	0,54*	0,18*	0,00	0,01	0,43*	0,56*	-0,28*	0,20*		
P					0,73*	0,24*	0,61*	0,25*	-0,15	-0,02	0,35*	0,33*	-0,19*	-0,06		
pH				0,63*	0,70*	0,23*	0,51*	0,23*	0,08	-0,07	0,37*	0,46*	-0,15	0,04		
Zn			0,43*	0,64*	0,36*	0,29*	0,35*	-0,03	0,20*	-0,27*	0,13	0,04	-0,16	-0,23*		
Fe		0,22*	0,24*	0,46*	0,35*	-0,06	0,51*	0,05	-0,12	-0,12	-0,35*	-0,27*	-0,67*	-0,45*		
Cu		0,80*	-0,05	0,16	0,23*	0,22*	-0,06	0,57*	0,28*	-0,15	-0,01	-0,29*	-0,14	-0,52*	-0,21*	
Areia	-0,04	0,23*	0,15	-0,15	-0,11	-0,22*	-0,32*	-0,44*	-0,88*	-0,02	-0,30*	-0,66*	-0,74*	-0,62*	-0,87*	
MO	0,11	-0,58*	-0,64*	-0,24*	-0,46*	-0,57*	-0,71*	0,00	-0,63*	-0,19*	-0,15	0,06	-0,04	-0,23*	0,50*	0,00
	Areia	Cu	Fe	Zn	pH	P	SB	K	Mn	Silte	Na	B	Ca	Mg	CIC	Argila

pH: potencial hidrogeniônico, P: fósforo, Cu: cobre, Zn: zinco, Fe: ferro, Mn: manganês, B: boro, MO: matéria orgânica, CIC: capacidade de troca catiônica, Ca: cálcio, Mg: magnésio, Na: sódio, K: potássio, SB: saturação de bases. *:Coeficiente de correlação estatisticamente significativo a um nível de $\alpha = 0,05$.

ANEXO

Figura 21 - Incremento da produção em toneladas de cana por hectare (THC, %), relacionado ao conteúdo de matéria orgânica expressada em porcentagem (MO, %), como resposta às aplicações de nitrogênio na cultura de cana-de-açúcar na Guatemala.



Fonte: CENGICAÑA, 2012.